Regulation of (generative) AI requires continuous oversight (AustLII submission on the 'Safe and responsible AI in Australia' Discussion Paper)

Graham Greenleaf, Andrew Mowbray and Philip Chung*

24 July 2023

Contents

* Graham Greenleaf is Professor of Law and Information Systems, UNSW Sydney; Andrew Mowbray is Professor of Law and Information Technology at UTS; Philip Chung is Associate Professor of Law at UNSW Sydney; Greenleaf and Mowbray are co-founders of the Australasian Legal Information Institute (AustLII). Chung is the Executive Director of AustLII. The authors' background relevant to this submission is in the Appendix. AustLII <http://www.austlii.edu.au> is Australia's leading provider of free access to legal information.

Australia's Department of Industry, Science and Resources has invited interested parties to make submissions on a Discussion Paper (DP) *Safe and Responsible AI in Australia*.[1] The DP lists 20 questions on which submissions are sought. The DP builds on a *Rapid response information report – Generative AI: Language models and multimodal foundation models* (RRI Report)[2] commissioned by the National Science and Technology Council (ANSTC), which contains the technical assumptions about generative AI that are used in the DP, as well as regulatory background.

This submission by researchers from the Australasian Legal Information Institute (AustLII) addresses the most important general issues identified in the Discussion Paper and suggests the best strategies to address them.

1. Submissions on issues in the Discussion Paper

1.1. Are threats or opportunities more important?

Adoption of AI systems in Australia is assumed to be relatively low (DP p.3), but we query whether this is so, particularly because of the apparent high rate of take-up of automated decision-making (ADM) systems in both the private and public sectors. Robodebt was seen as a wake-up call in Australia, but it may be that we have just been sleep-walking into broader use of AI (or systems with equivalent results) than we realised.[3] This question should be examined at the outset of determining an AI policy for Australia, and kept under regular review.

Irrespective of the rate of take-up, we submit that the main challenge for Australian policy is not to identify opportunities to capitalise on AI. The market will identify opportunities and firms willing to find investment funds. Regulation is only relevant to take-up of opportunities if there are impediments to implementation of AI in particular industries, or in the take-up of investment. In that case a body that seeks to identify these impediments and propose their removal where justified could be valuable. But pouring money into attempts to 'pick winners' is rarely successful.

Far more urgent is the need to identify those aspects of the development of automated systems (possibly involving AI including generative AI), particularly those that may pose considerable risk to Australia as a whole, or to particular segments of Australian industry or society.

We submit that the rate of take-up of automated decision-making (ADM) systems in both the private and public sectors in Australia, to identify those that may pose considerable risk to

---

[1] *Supporting responsible AI: discussion paper*, 1 June 2023 <https://consult.industry.gov.au/supporting-responsible-ai >.

[2] Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023, March 24). Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs). Australian Council of Learned Academies; see <https://www.chiefscientist.gov.au/GenerativeAI>.

[3] Mowbray, Andrew and Chung, Philip and Greenleaf, Graham, Applying the Rule of Law in Automated Decision Systems through Rules as Code (February 10, 2023). Submission to the Robodebt Royal Commission, 2023, UNSW Law Research Paper No. 23-4, <https://ssrn.com/abstract=4355989>.

Australia, should be examined at the outset of determining an AI policy for Australia, and kept under regular review.

1.2. Definitions: What is the correct focus of regulation?

The DP (pgs. 1-2) proceeds on the assumption that we should be focusing on the regulation of 'AI'. We have a different view, and *submit that regulation should be aimed at two things:*

(i) Regulation of the use of specific applications of underlying AI technologies; and

(ii) Regulation, by imposition of conditions on any use of a particular underlying AI technology, and therefore (for practical purposes) of their development.

Regulation should not aim to regulate the development of an underlying technology or its application per se, by preventing research into the technology or its application.

For example, we should not aim to prevent research into the technology of automated facial recognition (AFR), but might prohibit the use of applications of AFR at public events, or in commercial venues, and might require that any development of AFR underlying technology should comply with conditions designed to avoid unlawful discrimination.

The definitions used in the DP do not adequately distinguish between underlying AI technologies and the applications of these technologies. In respect of conversational AI, for example, the current approaches rely upon underlying large language models (LLM) based around machine learning and neural networks. In this context, textual generative AI is an application of these underlying technologies to create dialogues and documents. In other contexts, similar technologies can be used to create other types of artifacts such as artistic, musical and dramatic works. 'Generative AI' is not a separate type of AI but rather a particular application of AI.

1.2.1. Explicit and implicit programming

DP 1.2 provides a definition of 'AI', apparently based on ISO definitions, the essence of which is that 'AI' refers to 'systems that generate predictive outcomes … without explicit programming'. No such definition appears in the RRI Report, and we consider the definition grafted on by the DP is confusing and unhelpful.

Another criticism of the DP definition is that 'explicit programming' is not defined, and can mean a number of things, such as the distinction between procedural (explicit) and declarative (implicit) coding, a key difference between one older form of AI, 'rule-based systems' (implicit) and more traditional (explicit) programming. But the absence of explicit programming can also be used to refer to systems based on machine learning. So, the definition is not very precise. It is also not helpful because many systems we would consider to be 'AI', while they may be based on non-explicit programming, will also contain elements of explicit programming where that is efficient.

Just as important is that it may be possible to achieve by explicit programming the same result as is achieved by non-explicit programming, particularly once the non-explicit programming demonstrates that the result can be achieved. Where this is so, the use of 'AI' should make little difference to whether production use of a system is regulated, it should instead be the level of risk involved which determines the need to regulate. For example, a

trivial piece of declarative code may be 'AI' which can assist a user in deciding what to choose from a restaurant menu, but it is the fact that its use involves minimal/nil risk that means it should not be regulated.

Nevertheless, it is possible to accept that 'AI' provides a useful term for discussion of regulation, provided it is not too technologically restrictive, nor so broad as to encompass all programming. This could be achieved by altering the definition of 'AI' so that it refers to 'without explicit programming, or which achieves a similar result by other means'.

Whatever definition of AI is adopted, we submit that it should remain clear that regulation should apply to applications of underlying technologies, and in some cases the imposition of conditions on any use of those technologies (and thus of their development).

1.2.2. 'Hallucinations' or fabrications?

Concerning terminology, the DP and the RRR both refer to 'hallucinations' from generative AI when it produces 'entirely erroneous outputs' (DP p. 7). This is a common usage, but we submit that it a misleading usage which should be dropped. These 'hallucinations' include the fabrication of facts that don't exist, and the citation of journal articles or legal cases to support an argument that are either invented, or if they exist they do not support the proposition for which they are cited.[4] They would be more accurately described as 'deceptive and reckless mis-statements' or 'fabrications', either of which convey potential for liability that 'hallucination' does not.

1.3. Assessment of the most urgent dangers - 'Apocalypse now'?

During 2023 groups of experts have expressed very high levels of concern about the rapid development of generative AI, and more broadly, Artificial General Intelligence (AGI), and proposed some extreme reactions. In March 2023, the Future of Life Institute issued an open letter (with over 30,000 signatories) to "call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT4", and for governments to 'step in and institute a moratorium' if this was not observed.[5] On 30 May 2023, a group of over 350 extremely high profile AI scientists and other persons released a one sentence Statement on AI Risk: 'Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war'.[6] No governments have yet signed on. We can (and do) accept that AI regulation is a 'global priority' without necessarily endorsing sweeping measures such as a proposed moratorium on research that would stop before the year is out.

Australia should avoid being either an evangelist or a catastrophiser in relation to AI.[7] The measures that are needed must be continuous and detailed, not a 'one-off' broad solution, if only because we have as yet only very limited understanding of what needs regulation.

---

[4] For example, John Naughton 'A lawyer got ChatGPT to do his research, but he isn't AI's biggest fool' *The Guardian*, 4 June 2023 <https://www.theguardian.com/commentisfree/2023/jun/03/lawyer-chatgpt-research-avianca-statement-ai-risk-openai-deepmind>.

[5] 'Pause Giant AI Experiments' <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

[6] Centre for AI Safety 'Statement on AI Risk' <https://www.safe.ai/statement-on-ai-risk#open-letter>

[7] Matthew Knott 'Government may force companies to label AI content to prevent deep fakes; *SMH*, 16 June 2023, quoting Industry Minister Ed Husic.

The dangers posed by generative AI range from the very specific to the very general. We emphasise the following dangers, some of particular relevance to Australia:

- ***Costs and equity*** Large Language Models (LLM) used for generative AI are enormously expensive to run, with each generation of LLM becoming far more expensive than the previous. The costs arise from many factors: the currently high costs of GPU chips; the enormous consumption of electricity; and high-end hardware.[8] This cost factor poses very great risks for social equity, because it is possible that applications based on high quality LLMs may only become available to the wealthiest organisations and publishing houses, able to recover the costs through high profits, consultancy fees or transaction costs. Experiments in building lower-cost LLMs with a different and less costly architecture, and able to be run on 'consumer-quality' equipment, are underway but as yet inconclusive,[9] and need to be kept under observation.

- ***Fabrications*** ChatGPT's hallucination problem (which we call 'fabrications') might not be fixable, so it might only be able to be used safely in very limited situations involving low risk: [10] '… large language models get more accurate when they debate each other, but factual accuracy is not built into their capacity',[11] and other solutions have not been found. Systems like ChatGPT are a model of how people use language, but do not have an explicit or self-reflective deep understanding of how the world works, or an ability to use logic, mathematics and common sense, beyond what language usage indicates.[12] The dimensions of this problem are unknown.

- ***Inherited vulnerabilities*** Vulnerabilities in a foundation model are likely to be inherited in models derived from it.

- ***Irresponsible use*** Applications of LLM, like ChatGPT, are being used to influence decision-making in dangerous and inappropriate contexts. For example, the Australian Research Council has found it necessary to issue a *Policy on Use of Generative Artificial Intelligence in the ARC's grants programs* because of evidence that grant assessor's reports were being written with ChatGPT.[13]

- ***Content appropriation*** Models underlying generative AI could obtain unimpeded access to texts or other content, in order to generate new creative content, but with

---

[8] Will Oremus 'AI chatbots lose money every time you use them. That is a problem.' Washington Post 5 June 2023 <https://www.washingtonpost.com/technology/2023/06/05/chatgpt-hidden-cost-gpu-compute/>

[9] See, for example, Technology Innovation Institute's Falcon-7B model <https://huggingface.co/blog/falcon>.

[10] Gerrit De Vynck, 'ChatGPT 'hallucinates.' Some researchers worry it isn't fixable' Washington Post 30 May 2023 <https://www.washingtonpost.com/technology/2023/05/30/ai-chatbots-chatgpt-bard-trustworthy/ >

[11] Tate Ryan-Mosley, 'It's time to talk about the real AI risks' <https://www.technologyreview.com/2023/06/12/1074449/real-ai-risks/> *MIT Technol0gy Review* 12 June 2023

[12] Infocomm Media Development Authority (IMDA) / AICADIUM *Generative AI: Implications for Trust and Governance*, IMDA, Singapore, 2023, p.9.

[13] Donna Lu, 'Are Australian Research Council reports being written by ChatGPT?' *The Guardian* 8 July 2023 <https://www.theguardian.com/technology/2023/jul/08/australian-research-council-scrutiny-allegations-chatgpt-artifical-intelligence>

ruinous consequences for Australian copyright owners, who may find it very difficult to identify when their rights have been infringed or do anything about it. The difficulties will be multiplied if the data is exported overseas.

- ***Personal data theft*** If the personal data of Australian individuals becomes available (on a large scale) to developers of LLMs, then AI systems could result where it becomes impossible to trace how this personal data is being used, with serious consequences for the individual concerned or other individuals. Again, these problems will be multiplied if the personal data is exported overseas.

- ***Identity theft*** Other privacy dangers of systems supporting generative AI include a tendency to memorise sections of data records (which can be identifying), rather than just using data items; it may be possible to recreate training data by querying the model; and where a model incorporates user prompts to further train the model, users may find that their identifiable data has become part of the model.[14]

- ***Facial recognition technology additional dangers*** FRT can have sufficiently controlled and beneficial uses (for example, passport recognition at airports), but there are many uses of the technology which are already being recognised across the world as unjustifiably dangerous, and in breach of existing data privacy laws with resulting high penalties.[15] But if the facial data collected is transferred in bulk to a jurisdiction such as China, such existing remedies might prove to be too little. Whether the uses of facial recognition will proliferate out of control is not yet clear.

- ***Malicious code*** The ease with which persons without coding expertise can generate code could easily lead to a proliferation in the intentional generation and distribution of malicious code.[16]

These problems are not likely arise overnight, if they arise at all. They will occur incrementally, and what is important is that there should be an appropriate expert body in Australia which is continuously monitoring development, reaching an opinion on whether they are becoming dangerous, and informing the appropriate regulatory body, the government and the public, that this is the case.

The potential advantages of the use of AI, also ranging from the very specific to the very general. They make the use of generative AI difficult to resist, irrespective of risk:

- *Automated human tasks* Automation of some tasks now undertaken by humans (eg self-driving vehicles in some situations), particularly where human performance

---

[14] IMDA *op cit*, p.10.

[15] For example, Clearview AI, the most notorious facial recognition company, has been fined in numerous jurisdictions, and effectively put out of business, for processing images taken from the web without a valid legal basis: Garante (Italy) 10.02.2022 fined Clearview €20M; Hellenic DPA (Greece) 13.7.2022 fined Clearview AI Inc €20M; CNIL (France) fined Clearview AI Inc €20M; See Greenleaf, Graham 'Global Data Privacy Laws: EU Leads US and the Rest of the World in Enforcement by Penalties' (2023) 181 *Privacy Laws & Business International Report* 24-29 <https://ssrn.com/abstract=4409491>. In Australia, a similar result was reached by the Australian Information Commissioner and Privacy Commissioner that Clearview was in breach of the *Privacy Act*, although this is under appeal: see DP p.11.

[16] MDA *op cit*, p.11.

standards are not very high, and automated systems can be shown to exceed those standards.

- *Automated translation* Automated translation between languages has been transformed even further, including from human languages to code, and from poetry in one language (eg Urdu) to another (English).[17]

- *Automated form completion* An Indian developer trained GPT on all Indian government application documents, so the system could complete them, although in a different language. The whole population of India is able to benefit from this.[18] Microsoft's[19] and GitHub's 'copilot' programs are other examples.

1.4. Guiding principles for AI including generative AI applications

Since 2017 there has been a proliferation of sets of principles (variously named) focusing on ethical uses of AI.[20] Various authors have found that a 'striking … overlapping consensus … has emerged as to the norms that should govern AI.[21]

One set of early principles (2018), developed by civil society organisations, which deserves more attention is the *Universal Guidelines for Artificial Intelligence* (UGAI)[22] because it contains unusual consumer-oriented principles such as prohibitions on secret profiling and national scoring, and an obligation to terminate systems beyond human control. The UGAI have been endorsed by 300 experts and 60 associations. Another valuable source from a civil society perspective is the report on *Artificial Intelligence and Democratic Values* ('2022 AI Index') by the Center for AI and Digital Policy[23] which includes assessment of Australia's progress among 75 countries.

---

[17] Steven Levy 'Microsoft's Satya Nadella Is Betting Everything on AI' *Wired* 13 June 2023 <https://www.wired.com/story/microsofts-satya-nadella-is-betting-everything-on-ai/>

[18] ibid

[19] Wikipedia: Microsoft 365 Copilot

[20] For a summary, see Chesterman, Simon 'From Ethics to Law: Why, When, and How to Regulate AI' (April 29, 2023), forthcoming in *The Handbook of the Ethics of AI* Ed. David J. Gunkel (Edward Elgar Publishing Ltd.), NUS Law Working Paper No. 2023/014 <https://ssrn.com/abstract=4432941>

[21] Chesterman, ibid, citing Fjeld et al, 2020, Hagendorff, 2020 and Jobin et al, 2019; for a very critical perspective, see Clarke, Roger (2019) 'Principles and Business Processes for Responsible AI' *Computer Law & Security Review* 35, 4 (2019) 410-422, PrePrint at <http://www.rogerclarke.com/EC/AIP.html>, incl. 'The 50 Principles' <http://www.rogerclarke.com/EC/AIP.html#App1>;

[22] The Public Voice *Universal Guidelines for Artificial Intelligence*, Brussels, 2018.<https://archive.epic.org/international/AIGuidleinesDRAFT20180910.pdf>

[23] CAIDP.ORG *Artificial Intelligence and Democratic Values ('2022 AI Index')* 10 April 2023 <https://www.caidp.org/reports/aidv-2022/>

1.4.1. Existing Australian principles and requirements

The Australian government's eight *Artificial Intelligence (AI) Ethics Principles*[24], based on an IEEE publication, and consistent with the OECD's Principles on AI, are given by the DP (p. 14) as a good example of such a set of principles.

The other most relevant Australian statement of AI principles is the New South Wales (NSW) Government's *AI Assurance Framework* (NSW Framework), which assists NSW government agencies to design, build and use AI-enabled products and solutions, but also imposes obligations on them as to how they do so:[25]

> 'From March 2022, the AI Assurance Framework [is] required for all projects [within NSW government] which contain an AI component or utilise AI-driven tools. This includes the use of large language models and generative AI which are explicitly within scope of the application of the Assurance Framework. However a project is not expected to use the framework if:
>
> - It uses an AI system that is a widely available commercial application, and
> - The solution is not being customised in any way or being used other than intended.'

The NSW *AI Ethics Principles* are mandatory for NSW agencies (NSW Circular DCS-2020-04), and set out five broad ethical principles:[26]

- 'Community benefit – AI should deliver the best outcome for the citizen, and key insights into decision-making
- Fairness – Use of AI will include safeguards to manage data bias or data quality risks
- Privacy and security – AI will include the highest levels of assurance
- Transparency – review mechanisms will ensure citizens can question and challenge AI-based outcomes
- Accountability – decision-making remains the responsibility of organisations and individuals'

The NSW AI Assurance Framework[27] sets out in relation to these five broad principles, a complex set of checklists which agencies must complete and follow consequences depending on their answers. Consequences can include pausing the project (for example, until an impact assessment is done) or submitting the project for evaluation by a higher level review body. The consequences depend to a large extent on risk assessments, and complex tables of risk factors are provided.

The NSW approach is the most sophisticated yet developed in Australia, set out over 70 pages. It was developed before generative AI was well-known, but is considered to apply to it, and it is now being revised, including in light of generative AI.

---

[24] Dept. of Industry, Science and Resources *Australia's AI Ethics Principles* (2019) <https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles>

[25] Digital.NSW *NSW Artificial Intelligence Assurance Framework*, undated (prior to March 2022) <https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-artificial-intelligence-assurance-framework>

[26] *ibid*

[27] Application requirements, *NSW Artificial Intelligence Assurance Framework* <https://www.digital.nsw.gov.au/sites/default/files/2022-09/nsw-government-assurance-framework.pdf>

Chesterman finds that virtually all principles since 2018 include six themes,[28] which differ in some respects from the 'Australian Principles' and NSW Framework, but are also OECD-consistent.

1.4.2.Ten proposed more comprehensive principles

We submit that there should be a set of principles used to guide Australian regulation, that the principles should be based on international consensus, should be as consistent as possible across all Australian jurisdictions, and should be as comprehensive as needed. While we agree with the content of the DP's 'Australian principles' and the NSW principles and Framework, we consider that neither of them as clearly include some important aspects of the six themes identified by Chesterman (such as augmenting human abilities; explainability; and provision of remedies), neither set adequately reflects the need for protection of copyright. We have therefore amended the 'Australian principles' (new text is indicated by underlining), resulting in a more comprehensive but still succinct ten principles. *We submit that it would be preferable to adopt the following 'Ten guiding principles', which are a modification of the 'Australian principles'. The NSW Framework's proposed approach to implementation also needs to be taken into account.*

The first five principles are objectives (ends) that we wish to achieve by utilisation of AI. The last five principles are mechanisms (means) that we must use if we are to achieve these ends. Achievement of both ends and means are essential to the proper regulation of AI in Australia.

*1. **Human, societal and environmental sustainable benefit**: AI systems should benefit, or should not harm, individuals, society and the environment, and those benefits should remain sustainable given the availability of resources.*

*2. **Human-centred values**: AI systems should respect human rights, diversity, competition and the autonomy of individuals.*

*3. **Human control:** AI systems should augment rather than reduce human potential, and should remain under human control.* An obligation on developers to *terminate* the existence of a system which is likely to operate beyond human control (GP 3) is derived from the Universal Guidelines for AI.

*4. **Fairness:** AI systems should be inclusive and fair and should not involve or result in impermissible discrimination against individuals, communities or groups.*

*5. **Privacy protection:** AI systems should respect and uphold privacy rights and personal data protection.*

*6. **Reliability, safety and security**: AI systems should perform as intended, and be resistant to failure (whether accidental or from intentional interference) and to breaches of security.*

*7. **Transparency and explainability:** AI systems should have transparency, explainability and responsible disclosure so people can understand when they are being significantly impacted by AI, can find out when an AI system is engaging with them, and can obtain an explanation of actions or decisions affecting them.*

---

[28] Chesterman, op cit, p.3

8. ***Contestability:*** *When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.*

9. ***Accountability and remedies:*** *Persons (individual and corporate) responsible for the different phases of the AI system lifecycle should be identifiable, should be accountable for the outcomes of the AI systems, and should be liable to provide remedies for harmful impacts of the system.*

10. ***Copyright protection****: AI systems should protect copyright in the use or expression of data created or expressed by others, while making appropriate allowance for the public domain, for open source development of AI tools and applications, and defences/exceptions for use of evaluation tools.*

Each of these proposed changes in the Guiding Principles (GPs) deserves some detailed explanation, but their desirability will usually be apparent on their face. Some brief comments on the main changes:

- Benefits must be *sustainable* (GP 1). For example, generative AI systems require very high use, both in training and utilisation. Also, AI systems that require chips manufactured in Taiwan may develop considerable supply chain problems.[29]
- *Augmentation* of human abilities, preserving human control (GP 3) should generally be our aim in developing AI systems, not the complete replacement of human input.[30]
- An obligation on developers to *terminate* the existence of a system which is likely to operate beyond human control (GP 3) is derived from the Universal Guidelines for AI.
- Guarantees against system failures are unrealistic, but appropriate levels of *resistance to failure* (GP 6) (including data breaches) are essential.
- Development of AI systems incapable of explanation of their actions/conclusions is inherently high risk, so *explainability* (both in theory and practice) (GP 7) is of high value.[31]
- *Remedies* (GP 9) must usually be included, otherwise accountability is largely meaningless. 'Responsibility' without sanctions is usually ineffective.[32]
- Most generative AI systems rely on re-use of data created by others, so it is essential that *copyright* in data or expressions is protected (GP 10).

---

[29] Chris Miller, 'The Chips That Make Taiwan the Center of the World' (5 October 2022) *Time* <https://time.com/6219318/tsmc-taiwan-the-center-of-the-world/>.

[30] See Clarke *op cit*; see also Chesterman *op cit*.

[31] Mowbray, Andrew and Chung, Philip and Greenleaf, Graham 'Explainable AI (XAI) in Rules as Code (RaC): The DataLex approach' (2023) *Computer Law & Security Review*; pre-print <https://ssrn.com/abstract=4093026>

[32] Greenleaf, Graham 'Accountability Without Liability: 'To Whom' and 'With What Consequences'? (Questions for the 2019 OECD Privacy Guidelines Review)' (May 6, 2019). UNSW Law Research Paper No. 19-67, <https://ssrn.com/abstract=3384427 >

Do we need these ten Guiding Principles (or any similar set)? It is arguable that the most important sources of ethics/regulation of AI comes from 'the considerable value in using human rights law to evaluate and address the complex impacts of AI on society'.[33] In our view, they are needed because, although human rights laws may provide the substance of some of the GPs, there is a need for a succinct set of principles to identify the most relevant human rights laws, and in addition to identify other Principles needed to supplement them. The aim is not to rewrite human rights laws, but to make it easier to engage them where they are relevant.

1.5. A continuous oversight body

In our submission, *what Australia needs most to be able to safely and responsibly regulate AI (and particularly generative AI), is a continuous source of expert advice which will regularly report to existing regulatory bodies, to government and to the public, updating them on whether there are significant changes to our ability to uphold the principles (like those above) on which regulation of AI is based, and (if so) making proposals concerning regulatory changes that are needed.* Its tasks would not include the identification of commercial or public sector opportunities (which could easily distort the carrying out of its main task), but it could recommend desirable regulatory changes to remove impediments to these opportunities.

For purposes of discussion we will refer to the 'Australian Advisory Board on Regulation of AI' (the AI Board). *The AI Board would have a remit of two to three years, independence so that it could give frank advice, and an obligation to produce six monthly reports. The AI Board should preferably consist of ten members or fewer. Given Australia's federal structure, it will need to liaise with counterpart bodies in States and Territories concerning AI use by their public sectors.*

Taking into account the bodies already involved in work on AI regulation (see DP, Attachment AI), consideration should be given to including on the AI Board the following: the Chief Scientist; at least two AI experts, one from the National AI Centre; the eSafety Commissioner (or representative); the Privacy Commissioner (or representative); the ACCC, primarily in relation to consumer issues and digital platforms; the Australian Human Rights Commission, particularly in relation to discrimination issues; a copyright expert; at least one expert on the regulation of technology; and an expert on defence/security issues.

1.6. Australian consistency with, and input into, global AI regulation

Australia cannot isolate itself from global AI developments. It would be much better if international regulation (preferably global) resolved most issues before it was necessary for them to be regulated to accord with Australian standards. *As well as keeping abreast of these developments, Australia should aim to provide inputs to influence international developments where possible,* Australia needs to consider both short and long term perspectives on this.

---

[33] Raso, Filippo and Hilligoss, Hannah and Krishnamurthy, Vivek and Krishnamurthy, Vivek and Bavitz, Christopher and Kim, Levin Yerin, 'Artificial Intelligence & Human Rights: Opportunities & Risks' (September 25, 2018). Berkman Klein Center Research Publication No. 2018-6, <https://ssrn.com/abstract=3259344>

1.6.1.Short-term aims

In the **short term**, *the most advanced international source of AI regulation is likely to be the European Union*, whose nearly-completed AI Act takes a risk-based approach, categorising risks as Minimal, Limited, High or Unacceptable, with corresponding obligations on both providers and users (see DP p.17 and Appendix B). The EU's AI Act will have inherent merits as a piece of legislation which has been carefully thought out and debated across all EU institutions since at least 2021. The EU's approach is likely to be emulated by countries outside Europe (including Australia), part of the so-called 'Brussels effect' of EU standard-setting.[34] The international norm-setting effects of EU laws are such that many businesses will be more comfortable with non-EU laws that emulated norms found in the EU. Stanford researchers suggest that 'the EU AI Act is the most important regulatory initiative on AI in the world today' and that 'Policymakers across the globe are already drawing inspiration from the AI Act, and multinational companies may change their global practices to maintain a single AI development process.'[35]

Utilising the European Parliament's final draft of the EU AI Act, these researchers evaluated 'whether [10] major foundation model providers currently comply with these draft requirements and find that they largely do not. Foundation model providers rarely disclose adequate information regarding the data, compute, and deployment of their models as well as the key characteristics of the models themselves. In particular, foundation model providers generally do not comply with draft requirements to describe the use of copyrighted training data, the hardware used and emissions produced in training, and how they evaluate and test models.'[36] However, they consider that their assessment shows that 'it is currently feasible for foundation model providers to comply with the AI Act.' They conclude that policymakers globally should prioritize transparency, 'informed by the AI Act's requirements', and assess how the 10 foundation models considered meet 12 criteria focussing on transparency.

There is good sense in keeping Australia's AI regulation as consistent as possible with the approach adopted by the EU. Australia's AI Board could consider the evidence used by the EU in classifying the practices identified under each of the four categories of risk and decide whether or not to recommend similar regulatory requirements in Australia. The priorities for the order of assessment would depend on the Australian environment.

The other most prominent short-term initiative requiring consideration is the US Biden Administration's 'Voluntary Commitments' from seven leading AI companies, announced on

---

[34] Greenleaf, Graham 'The 'Brussels Effect' of the EU's 'AI Act' on Data Privacy Outside Europe' (2021) 171 *Privacy Laws & Business International Report* 1, 3-7, <https://ssrn.com/abstract=3898904>; see also Wikipedia: Brussels Effect < https://en.wikipedia.org/wiki/Brussels_effect>

[35] Rishi Bommasani, Kevin Klyman, Daniel Zhang and Percy Liang 'Do Foundation Model Providers Comply with the EU AI Act?' Stanford University *Center for Research on Foundation Models*, June 2023 <https://crfm.stanford.edu/2023/06/15/eu-ai-act.html>

[36] Bommasani et al, ibid

21 July 2023.[37] The announcement says the companies make eight commitments in relation to AI: to do pre-release security testing; to share (unspecified) information on risk management; to protect proprietary information about models (particularly model weights); to facilitate third party discovery and reporting of vulnerabilities; to ensure users know when content is AI-generated (eg watermarking); to publicly report AI systems' capabilities/ limitations; to 'priorities research on … societal risks'; and 'to help address society's greatest challenges'. How the commitments are made is not specified. Australia is one of 20 countries said to have been 'consulted'.

These commitments generally lack any concrete obligation to do anything of substance, except perhaps in relation to some aspects of transparency, and with no consequences for failure to do so. Some of the commitments are to do things that would be expected of such companies. Missing are substantive obligations in areas such as privacy and copyright protection. Commitment to watermarking AI-generated content can be seen as a smokescreen for not addressing whether the content used for such generation is legitimately used. Australia should not regard this US initiative as a significant guide to what should be done here.

1.6.2.Longer-term aims

In the **longer term**, it is possible that there may be steps toward a binding international agreement concerning AI (or generative AI at least), if international opinion concludes that it involves extreme dangers (as some have warned). Treaties already exist concerning the extreme dangers of nuclear proliferation, and of some types of environmental hazards. UN Secretary-General Guterres has backed a proposal by some AI executives for the creation of an international AI watchdog body like the International Atomic Energy Agency (IAEA), while stressing that only UN member states can create it.[38] In July 2023 the UK Foreign Secretary will also convene the first briefing of the UN Security Council on the opportunities and risks of AI for international peace and security. However, the UN has as yet had no impact.

The OECD, as the broadest forum for industrialised Western-oriented countries, developed a set of AI principles which have been influential in other forums such as the G20, perhaps will be in the G7. A treaty could possibly emerge from that direction.

One of the most advanced developments of a draft international treaty on AI is the Council of Europe (CoE) Committee on Artificial Intelligence draft Convention.[39] It is a full draft Convention of 38 articles, and (as with other CoE 'open' conventions) it is possible for countries that are non-members of the Council (like Australia) to be invited to accede to the

---

[37] White House Briefing Room 'FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI', 21 July 2023 <https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/>

[38] Michelle Nichols 'UN chief backs idea of global AI watchdog like nuclear agency' Reuters 13 June <https://www.reuters.com/technology/un-chief-backs-idea-global-ai-watchdog-like-nuclear-agency-2023-06-12/ >

[39] Council of Europe Committee on Artificial Intelligence draft Convention *Zero Draft [Framework] Convention On Artificial Intelligence, Human Rights, Democracy And The Rule Of Law*, Strasbourg, 6 January 2023 <https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f>

treaty (as has occurred with Data Protection Convention 108, and with the Cybercrime Convention). Australia should give this potential treaty serious consideration.

On 8 June 2023, the United States and the United Kingdom announced the 'Atlantic Declaration for a Twenty-First Century U.S.-UK Economic Partnership',[40] in which the US welcomed the UK's 'plans to launch the first Global Summit on AI Safety, to be hosted in the United Kingdom this year, and commits to attend at a high level.' It welcomed 'ongoing activity internationally including at the OECD, UN, Global Partnership for AI, Council of Europe, and International Standards Organisations, as well as the G7 Hiroshima AI Process'. The UK effort, it said, will 'bring together key countries, as well as leading technology companies and researchers, to drive targeted, rapid international action focused on … exploring safety measures to evaluate and monitor risks from AI'. The UK would clearly like some key global role in AI, and is investing heavily in some aspects, but it is easy to over-rate its influence.

Other than the Council of Europe initiative, none of this yet sounds like the negotiation of an international treaty, but Australia must ensure that it be represented in these forums, including both the UK and UN ones, and preferably the CoE, so as to have input into what eventually emerges.

1.7. A risk-based approach to regulation

We submit that, as many others are advocating, *a risk-based approach to regulation of AI should be adopted and should be implemented in Australia by Commonwealth legislation and regulations insofar as the Constitution permits, and through an approach which emphasises bringing only the most dangerous applications within the regulatory structure in the first instance.*

In more detail, this would involve steps such as the following:

- A risk-based approach should be mandated through regulation, and should apply to both (federal) *public or private organisations*, and to *both developers and deployers* (users). States and Territories could adopt similar legislation applying to their public sectors, but this submission is not directed toward that.
- The *risk-based regulatory framework* would preferably comprise four levels of risk similar to the EU AI Act (Unacceptable, High, Limited, and Minimal) (DP Attachment B), but with 'Limited' renamed 'Moderate'. This is preferable to the less flexible three levels (Low, Medium, High) suggested in DP Box 4.
- Each level of risk would be accompanied by regulatory requirements imposed on both Developers (or Providers) of the AI system for use by others, and on Deployers (Users) of the AI system.
- The AI Board would recommend to government which level of risk should apply to particular Models/Applications it thinks should be brought within the Framework. It could consider the evidence used by the EU in classifying the practices identified

[40] The White House, Statement 'The Atlantic Declaration: A Framework for a Twenty-First Century U.S.-UK Economic Partnership' 8 June 2023 <https://www.whitehouse.gov/briefing-room/statements-releases/2023/06/08/the-atlantic-declaration-a-framework-for-a-twenty-first-century-u-s-uk-economic-partnership/ >

under each of the four categories of risk and decide whether or not to recommend similar regulatory requirements in Australia. The priorities for the order of assessment would depend on the Australian environment.

- The *inclusion of models or applications in categories of AI systems under the Framework* would, except in the (probably) few cases listed in the original legislation, be included by regulations or by subsequent legislation (depending on significance), ideally following advice to government by the AI Board.

- Systems in the *High Risk category* should be *required to comply with all of the GPs*.

  Systems in the *Moderate Risk category* should be required to comply with specified GPs, which may differ between Models/Applications. Applications recommended by the Board to be in particular categories should be able to provide (to government) a high level of justification for exceptions to particular GPs applying to them.

Systems in the '*Minimal risk*' category would have no mandatory GP obligations (except in relation to transparency), but voluntary adoption of the GPs would be encouraged.

1.8. A possible initial framework for Australian regulation

In our view, *the federal government should initially enact an 'AI Framework Act'*, including at least the following elements:

- It should *create an Australian AI Board,* similar to that discussed above in part 1.5, with sufficient resources to fund a small secretariat, and powers to compulsorily gather information relevant to AI.

- It should require assessment by government, within six months, of the rate of take-up of automated decision-making (ADM) systems in both the private and public sectors, and the rate of take-up of other AI systems in Australia.

- It should include the *Ten Guiding Principles for (Generative) AI* set out above in part 1.4.2. These are 'Guiding Principles' (GP), not mandatory requirements, which means in our view that, where regulation of AI systems is considered necessary, each of the 10 GPs should be adhered to in that regulation, except where there is considered to be justification for an exception. GP should therefore function like default requirement, where exceptions require justification.

- It should *implement a risk-based approach* to regulation, as discussed above in part 1.7.

- *It should make transparency mandatory* for all AI applications impacting upon Australian individuals and organisations, including the foundation models on which they are based. There are two aspects of this transparency. There should be mandatory reporting of the use of AI for High or Medium Risk applications, analogous to reporting under the Modern Slavery legislation. There should also be mandatory labelling of AI-generated content.

2. Summary of submissions

We submit that:

1. The rate of take-up of automated decision-making (ADM) systems in both the private and public sectors in Australia, to identify those that may pose considerable risk to Australia, should be examined at the outset of determining an AI policy for Australia, and kept under regular review.

2. Regulation should not be aimed at 'AI generally', but should be aimed at two things:
(i) Regulation of the *use of specific applications* of underlying AI technologies; and
(ii) Regulation, by *imposition of conditions on any use of a particular underlying AI technology*, and therefore (for practical purposes) of their development;
but should not aim to prevent research into the technology or its application

3. The definition of 'AI' should be altered so that it refers to 'without explicit programming, or which achieves a similar result by other means'.

4. 'Hallucinations should be dropped, and 'deceptive and reckless mis-statements' or 'fabrications' used instead.

5. There should be a set of principles used to guide Australian regulation, that the principles should be based on international consensus, should be as consistent as possible across all Australian jurisdictions, and should be as comprehensive as needed.

6. It would be preferable to adopt the 'Ten guiding principles' set out in part 1.4.2, which are a modification of the 'Australian principles'. The NSW Framework's proposed approach to implementation also needs to be taken into account.

7. For Australia to be able to safely and responsibly regulate AI (and particularly generative AI), there needs to be a continuous source of expert advice which will regularly report to existing regulatory bodies, to government and to the public, updating them on whether there are significant changes to our ability to uphold the principles on which regulation of AI is based, and make proposals concerning changes needed.

8. An 'Australian Advisory Board on Regulation of AI' (the AI Board) would have a remit of two to three years, independence so that it could give frank advice, and an obligation to produce six monthly reports. It should preferably consist of ten members or fewer.

9. As well as keeping abreast of international developments in AI regulation, Australia should aim to provide inputs to influence international developments where possible. In the short term, the most advanced international source of AI regulation is likely to be the European Union. In the longer term, it is possible that there may be steps toward a binding international agreement concerning AI (or generative AI at least).

10. A risk-based approach to regulation of AI should be adopted and should be implemented in Australia by Commonwealth legislation and regulations insofar as the

Constitution permits, as set out in part 1.7. It should involve bringing only the most dangerous applications within the regulatory structure in the first instance.

11. The Commonwealth's initial 'AI Framework Act' should involve at least the following:

    a. Create an Australian AI Board, similar to that discussed in part 1.5.

    b. Require assessment by government, within six months, of the take-up of AI in Australia.

    c. Include the *Ten Guiding Principles for (Generative) AI* set out in part 1.4.2.

    d. Implement a risk-based approach to regulation, as discussed above in part 1.7.

    e. Make transparency mandatory for all AI applications impacting upon Australian individuals and organisations, including the foundation models on which they are based.

3. Authors' qualifications

Graham Greenleaf AM is Professor of Law and Information Systems at UNSW Sydney; Andrew Mowbray is Professor of Law and Information Technology at UTS; Philip Chung is Associate Professor of Law at UNSW Sydney.

In 1995 Greenleaf and Mowbray were co-founders of the Australasian Legal Information Institute (AustLII – http://www.austlii.edu.au), which became a pioneer of the global Free Access to Law Movement (FALM - http://www.falm.info/). In 1996 they were joined by Chung, who is now the Executive Director of AustLII. AustLII is the dominant provider of comprehensive legal research infrastructure in Australia.

More directly relevant to this submission, Mowbray and Greenleaf were global pioneers in the development of AI applications applied to law, from the mid-1980s onward. They were joined in this work from 1995 by Chung, when it became part of AustLII's suite of software. Their AI work was under the name 'The DataLex Project', and its history from 1984-2001 is at <http://datalex.org>.[41]

Since the revival of interest in AI development from around 2015, the authors' main publications in the field have focused on building decision support systems using the DataLex tools,[42, 43]'explainable AI'[44], automated decision systems,[45] and 'scaling up' rules as code.[46]

Greenleaf also has an extensive track record of research and publications in relation to regulation of information technology, particularly in relation to data privacy,[47] and in relation to the copyright public domain.[48]

---

[41] Greenleaf, Graham and Mowbray, Andrew and Chung, Philip 'The Datalex Project: History and Bibliography' (January 3, 2018). UNSW Law Research Paper No. 18-4, <https://ssrn.com/abstract=3095897 >

[42] Greenleaf, Graham and Chung, Philip and Mowbray, Andrew, 'Building Datalex Decision Support Systems: A Tutorial on Rule-Based Reasoning in Law' (September 4, 2017). UNSW Law Research Paper No. 17-68, <https://ssrn.com/abstract=3034430 >

[43] Mowbray, Andrew and Greenleaf, Graham and Chung, Philip, Law as Code: Introducing AustLII's DataLex AI (November 16, 2021). UNSW Law Research Paper No. 21-81, <https://ssrn.com/abstract=3971919 >

[44] Mowbray, Andrew and Chung, Philip and Greenleaf, Graham 'Explainable AI (XAI) in Rules as Code (RaC): The DataLex approach' (2023) *Computer Law & Security Review* ; <https://ssrn.com/abstract=4093026 >

[45] Mowbray, Andrew and Chung, Philip and Greenleaf, Graham, Applying the Rule of Law in Automated Decision Systems through Rules as Code (February 10, 2023). Submission to the Robodebt Royal Commission, 2023, UNSW Law Research Paper No. 23-4, <https://ssrn.com/abstract=4355989>

[46] Mowbray, Andrew and Chung, Philip and Greenleaf, Graham, 'Representing Legislative Rules as Code: Reducing the Problems of 'Scaling Up'' (2023) *Computer Law & Security Review;* preprint <https://ssrn.com/abstract=4148039>

[47] See in particular Greenleaf, Graham *Asian Data Privacy Laws: Trade and Human Rights Perspectives* (Oxford University Press, 2014)

[48] See in particular Greenleaf, Graham and Lindsay, David *Copyright's Public Domains* (Cambridge University Press, 2018)