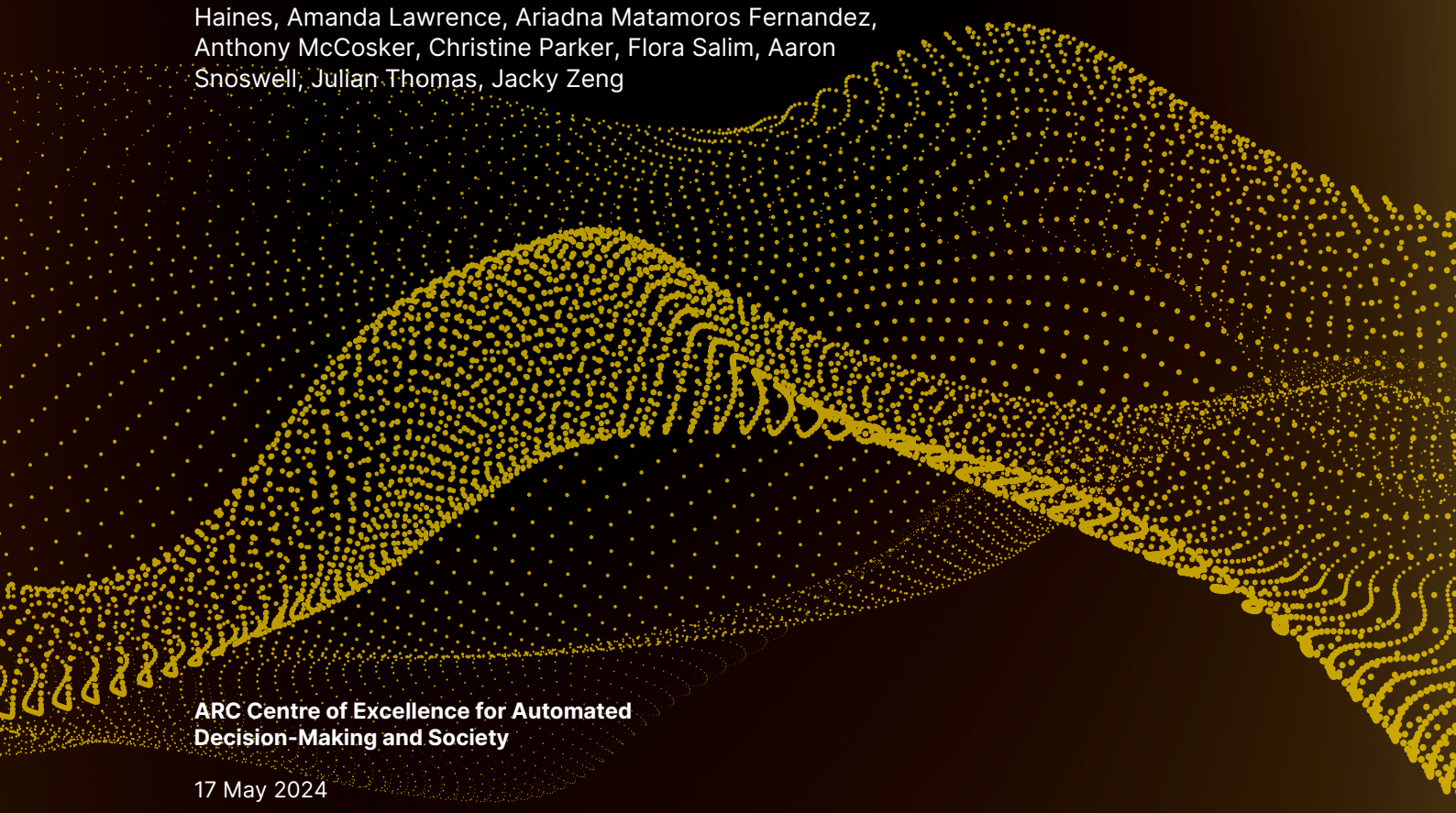
ADM+S

ARC Centre of Excellence for **Automated Decision-Making and Society**

ADM+S Submission to the Senate Select Committee on Adopting Artificial Intelligence

Lead author
Kimberlee Weatherall

Contributing authors
Dan Angus, Jose-Miguel Bello y Villarino, Jean Burgess, Henry Fraser, Jake Goldenfein, Melissa Gregg, Fiona Haines, Amanda Lawrence, Ariadna Matamoros Fernandez, Anthony McCosker, Christine Parker, Flora Salim, Aaron Snoswell, Julian Thomas, Jacky Zeng



ARC Centre of Excellence for Automated Decision-Making and Society

17 May 2024

About ADM+S

The ADM+S is pleased to have this opportunity to engage with this inquiry into the opportunities and impacts for Australia arising out of the uptake of AI technologies in Australia. The ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S) is a cross-disciplinary, national research centre established and supported by the Australian Research Council to create the knowledge and strategies necessary for responsible, ethical, and inclusive automated decision-making (ADM). This submission draws on research from across ADM+S, and on our August 2023 submission on *Safe and Responsible AI*. We refer the Committee to that submission for more extended discussion on many of the points below.

This submission

This submission is the product of a collaborative process involving direct contributions from the above researchers from ADM+S, as led and consolidated by Professor Kimberlee Weatherall (University of Sydney Law School). ADM+S researchers come from many different institutions, disciplines and perspectives. It should not be assumed that every contributing author, or every member of the Centre subscribes to every comment or recommendation in this submission. The submission represents our best effort to consolidate research and thinking in a way that can be useful to the Committee. For a further list of ADM+S research relevant to the Terms of Reference, see **Appendix B**.

Table of Contents

Recent trends and opportunities in the development and adoption of AI (especially generative AI)

Writing about trends in the development and adoption of AI is challenging in an environment where announcements of major developments are frequent. As this submission was being written, Google made a slew of significant AI-related announcements,[1] and OpenAI published video demonstrations of OpenAI's latest foundation model, which it is calling GPT-4o (omni).[2]

GPT-4o is multimodal, accepting input from any combination of text, audio, image, and video, and generating any combination of text, audio, and image outputs. Capability-wise, it appears to match or exceed other state-of-the-art foundation models across the board. For example, in automatic speech recognition (ASR), it beats OpenAI's own specialised ASR tool, Whisper.[3] Demonstration videos suggest that GPT-4o is particularly striking for its **realism in emulating human interaction**. Three key aspects contribute to this:

1. Real-time visual input capability: i.e. the model accepts visual input from a device camera in real time, and so appears to 'see' and 'understand' the world;

2. Significant improvements to the output audio (voice), including subtle sighs, tones and inflections, laughing and singing that make the speech sound more human than previous voice assistants like Siri or Alexa; and

3. Apparently improved latency, meaning the delay on response approaches human response times in conversation.

As with other recent releases, OpenAI has not published a formal research paper; rather, they have provided model documentation via a polished webpage, which is more accessible, but less informative. The release is positioned as a ready-to-use consumer product that is free to all users, rather than a scientific advance open to interrogation. The release follows closely behind Meta's move to incorporate a single AI assistant across all their consumer platforms (WhatsApp, Facebook, Instagram)[4] and Microsoft's ongoing incorporation of AI (including via the Copilot assistant) into its existing products and platforms.[5] It also coincides with the public launch of Google's Project Astra, a universal AI agent that uses real-time input to help in a range of daily tasks, including translation, productivity assistance, and learning.[6] This combination of near-simultaneous events signals a major impending shift in how everyday Australians will encounter, experience, and use AI:

- A shift from AI as a research-driven computer science field, to a multi-billion-dollar industry that engages directly with consumers. Put another way, we are seeing a shift to AI as consumer product.

- AI's increasing ubiquity across the digital media and information environment, workplace environments, and beyond, and the inevitability of digitally-connected Australians interacting with AI, not necessarily as a result of conscious decisions or choices, but as a result of AI being incorporated into the platforms and devices that are an essential part of daily life.

Ubiquitous AI is not necessarily well-understood AI. Not only are the implications of ubiquitous AI largely unknown and unexplored to date, but to make a more basic point: just because we are interacting with AI systems does not mean we understand them or know even at a high level how they work.

There is much we do not know about the AI systems we presently interact with. Transparency is a significant issue. One area of ADM+S research relates to the use of AI and ADM in news and media, including by social media platforms. ADM+S researchers have examined how AI is being used in search, recommendation, content moderation, and advertising,[7] including via two projects that use data donation to observe and analyse how people experience search, and what advertisements they see. The novel methods used in this project are necessary because it remains very difficult for independent researchers to understand how AI is impacting on our information and communications environment. The efforts ADM+S researchers have had to undertake, as well as the inadequacy of transparency tools provided by the platforms are set out in our Ad Observatory Technical and Data Report.[8] We have argued for better access to data for researchers, and appropriate transparency from providers, in order to ensure that Australians and Australian policymakers can receive independent, expert advice in this area. Development of the NCRIS/ARDC-funded Australian Internet Observatory[9] commencing in July 2024 will extend the work of ADM+S researchers to date, with the aim of significantly enhancing the observability of AI, ADM and digital platforms over the next four years. Future investment in social science research infrastructure to support the development of critical datasets and technical frameworks to support these endeavors is highly recommended.[10]

MFM-based agents such as GPT-4o can perceive and reason about the world in real time, owing to their ability to work with multimodal input through image, video, text, and other modalities. In the near future, given access to sensor and wearable data (e.g. from smartwatches), earables (e.g. Airpods), and eyewear computing (e.g. Meta Raybans), MFM agents may also have embodied representations of human users. A recent indication of these coming developments is Apple's Foundation Models for monitoring wellness and medical conditions, trained via thousands of Apple Watch users.[11]

Further, as shown in the recent demo of GPT-4o and Google's Astra, agent-to-agent communication and collaboration is now a reality. This leads to a much more complex world of human-multi agent interactions, possibly in a more serendipitous manner, and at scale. It opens new realms of possibilities of agencies and creativity, but also poses more risks.

Other new/emerging capacities: ADM+S has previously drawn attention to things that can be done with newer forms of AI and automation that might have been harder, or impossible, to achieve at scale before. These new capabilities warrant broader discussion about what is, and is not, acceptable/consistent with a flourishing Australian society and economy. We draw the Committee's attention to:

- **Automated, remote mass surveillance** and facial/biometric recognition (by public and private sector actors): there are strong public and private incentives to increase the identifiability of people in the public sphere. This may be useful (or even necessary) for law enforcement and security, to enforce norms and

conditions (such as self-exclusion of people with a gambling problem from venues; exclusion of disruptive people from sporting events); or to enable retailers to personalise or conduct longitudinal market research on their offerings. But there are also costs, and the current spread of these systems in public places in Australia warrants a public discussion regarding the extent to which, and how we want to protect privacy and anonymity.

- **'Social scoring**': linking data across domains for analysis and action (within the private and public sectors) such that a person's history and behavior in one area can impact their opportunities and activities in unrelated domains to an extent never before possible. What are the limits on the use of data across domains?

- **Invasion of mental privacy/autonomy:** we may not yet have 'mind-reading robots' or 'AI-facilitated behavioral manipulation', but there are strong incentives to develop technology to identify, predict, and influence individuals' mental states, intentions and actions. Should there be limits on private and public sector actors' ability to seek to record or use people's emotional states or intentions? When GPT-4o asks or seeks to infer how a user feels,[12] when does that cross a line?

Automation and increasing digital inequality: the imperative for efficiency across both public and private sectors drives increased automation; generative AI is expected to accelerate this trend. Informed engagement with automation should take into account impacts on people, especially (but not exclusively) vulnerable, disadvantaged or isolated groups who rely on income support or other forms of assistance in their daily lives. Harms include the intended and unintended impacts of system changes driven by automation, the inability to access or effectively use critical automated services such as banking or Centrelink, and the social costs of increasing digital inequality. ADM+S' Australian Digital Inclusion Index and Mapping the Digital Gap research tracks the levels of digital exclusion and its geographic and social distribution[13].

Increasing costs. While Generative AI enables many tasks to be done more cheaply and efficiently than before, the costs of providing services such as general use chatbots are also higher than those involved in longstanding web-based utilities such as search engines. Recent reports suggest a trend towards tiered products in this area, with Google Search considering a subscription model for Generative AI-enhanced Search.[14] One consequence of such a development would be new differentials in the affordability of important web services hitherto wholly supported by advertising, with economic consequences for both organisations and consumers.

The **institutional** impacts of automation also demand attention. ADM+S has also conducted research in partnership with the NSW Ombudsman, mapping the **use of automated decision-making by governments in NSW**, both state government departments and agencies, and local governments.[15] The research resulted in a published compendium of automated decision-making systems across NSW state and local governments.[16] The ADM+S report made a number of findings about AI (and ADM) use in government:

- ADM (which is not the same thing as AI, but can be facilitated by AI) is used across every NSW state government portfolio, and at the local government level;
- As of 2023, there was considerable interest in, but limited use of, **generative AI** in government. Departments and agencies identified potential cases where they were considering incorporating predictive analytics into existing structured decision-making processes;[17]
- Government departments and agencies were also confronting the question of whether or how to use or evaluate AI features being incorporated by existing providers into existing systems (a trend that has accelerated since);
- There were very different levels of AI literacy and governance systems across different parts of the NSW government; it was not always clear who had responsibility or knowledge of systems within a department or agency; and
- From a regulatory perspective, and a transparency perspective, it is not straightforward to identify where an 'ADM system' begins and ends.

Risks and harms

As discussed in our earlier submission to DISR,[18] the question is what **new or changed** risks of harm arise with the increased power and integration of AI, including generative AI. Different kinds of AI technology, and different ways of using the technology, give rise to different risks of harm,[19] and at varying scales or orders of magnitude. Some newer generations of AI are giving rise to new harms, and/or further amplifying existing harms. We need to be paying attention to **both** well-known and emerging risks. A (non-comprehensive) diagram that seeks to illustrate the range of risks and harms is provided below in **Appendix A**.

Certain risks are reasonably well known and have attracted research, and policymaker attention.

1. **Bias, unfairness, opacity, unpredictability:** Certain risks of harm are very well known and studied in the literature. Despite being well known, they have not been solved, and addressing them remains an important task. AI draws inferences from patterns in existing data.[20] When biases are embedded in the data used to train models, models tend to perpetuate those biases: this can lead to **discrimination** and more generally **unfairness**. Given a sufficiently complex model drawing on sufficiently large data, the basis on which such a system draws inferences may be **opaque** (interfering, in some cases, with people's rights to understand the reasons for a decision)[21] and sources and likelihood of error may be **unpredictable** (which again can interfere with people's rights, or interests, in sound, well-based decisions).

2. **Quicker, cheaper content generation amplifying communication harms:** Generative AI[22] enables cheap, rapid generation of content at scale, amplifying a range of potential communication harms: online abusive content; scams and fraud; mis/disinformation; and violence and extremist content. As the technology improves, and synthetic content becomes more convincing, it may become harder to differentiate authentic information and content with high integrity.[23] Research published by Anthropic also suggests that models are becoming measurably more persuasive over time.[24] These developments challenge traditional approaches that seek to promote online safety through a combination of regulatory, technical and media literacy approaches.[25]

3. **Concentration of power and data:** there is reason for concern that the most advanced forms of technology are being developed, and held, by a very small number of global actors; competition authorities across the UK, US and Europe are conducting a number of investigations into OpenAI, Microsoft, and generative AI investments more generally.

Other risks are perhaps less well-studied, less well-understood and warrant attention:

1. **Technology exemplified by GPT-4o enables more 'human-style' interactions**. This has clear benefits in expanding access to and use of technology. But we also need to better understand the (positive and negative) impacts of human-computer interactions that are closer to human-human interactions. Do people trust more readily (creating opportunities for manipulation or deception)? Do they reveal more information (creating privacy risks); and what are the implications of large-scale collection, and analysis, of human-computer dialogue? Are people more vulnerable (risking emotional harm) when dealing with such technology? What risks of exclusion or discrimination arise if such technology interacts with some people more readily than others, or differently, and what are the costs and benefits of technology that can interact with different people differently?

2. **Autonomy:** larger general-purpose systems can undertake more complex tasks, and chain tasks, in a more autonomous fashion, relying on predictions and inferences. This raises questions regarding how such 'agentic' AI can best be controlled/aligned to achieve (legitimate) human goals.

3. **Value/supply chains;** AI value chains are complex, giving rise to value/supply chain risks and problems in both identifying and mitigating risks, and allocating responsibility. Generative AI systems for example may involve incorporating procured datasets and pre-trained models, from different sources into single systems, leading to a diffusion of responsibility.[26] Existing laws have little capacity to manage dataset economies and the specific harms they generate, such as using personal data and copyrighted material to generate value that is unfairly distributed.[27]

4. **Emerging capabilities from very large models:** the largest state of the art models (sometimes referred to as 'frontier models') have demonstrated capabilities that

have surprised even experts in the field. Where capabilities are surprising, it can become hard to predict the risks posed by release or use of a model or system. In addition, some of these capabilities may give rise to concern: for example, evidence that in some cases systems have offered deceptive responses to human users.[28] These developments create research, regulatory, and security challenges.

Note that we address the issue of environmental impacts below.

These changes highlight **gaps and/or uncertainties in existing laws.** ADM+S listed some of these in our earlier submission.[29] It is worth noting that many cases or investigations initiated by regulators overseas to address harms emerging from the use of AI have depended on legal provisions that Australian law currently lacks (such as certain consumer and privacy rights). In many cases, the adoption of similar rules has been proposed in Australia for some time, but actual reform has been extremely slow, and lags behind equivalent nations.[30]

Shifts in AI technology have also created potential gaps in liability and **enforcement**.[31] For example, supply chain complexities require us to rethink how we assign responsibility. If it is hard to ascribe responsibility for specific 'AI errors', could fault inquiries focus on non-technical choices in design, documentation, risk management, deployment, marketing etc. that contribute to risks of harm?[32] Do we need to expand the concept of a 'manufacturer' for the purposes of product liability, and expand product liability to apply to software updates that lead to defects? Or perhaps we need to rethink consumer guarantees: at present, consumer guarantees seem likely to apply to downstream app providers, but not upstream foundation model providers.

The capacity of AI to cause harm on a scale and at a speed not previously possible is a strong argument in favour of supplementing current individually focused, *ex post* enforcement through court proceedings for the redress of harm, with a range of other legal mechanisms. There is **emerging international consensus on the need for risk-based regulation with *ex ante* elements:** requiring those designing, developing, and deploying AI to mitigate risks early on during design, development and deployment, and on an ongoing basis through monitoring and evaluation. We note that the government has committed to a risk-based approach to AI Regulation.[33]

The diffused effects of some AI systems – causing widespread, but low-level harm – may also demand adjustment in enforcement: for example collective enforcement mechanisms (such as class actions or granting rights of action to representative groups). Conceptualisations of harm that heavily draw from individualised criminal law frameworks tend to overlook social harms and might be incomplete to address all online harms.[34]

We also need to move our conversation about the risks of AI from a concern with discrete issues at the level of AI technology development and deployment and towards a more holistic view. Bias is a real and important risk from AI systems, but risk mitigation has so far tended to focus on technical solutions and de-biasing toolkits. Merely trying to control bias at the level of output has not yet proven effective, and even the largest technology companies struggle to deal with it credibly, as Google's recent embarrassing experience with its Gemini image generator showed.[35] Likewise, the possibility of aberrant and unpredictable outputs

from AI systems is a serious risk across a range of contexts, from chatbots to embodied robots and autonomous cars. But focusing regulatory and safety efforts on error-prevention will not be enough.[36]

Engineers in safety-critical fields like aviation take a 'systems' approach to risks, recognising that an error-free system is not necessarily a safe system. System safety practitioners understand that organisational, communicative, contextual, cultural and even aesthetic factors contribute to risks, and that effective regulation and harm prevention addresses socio-technical as well as technical dimensions of accidents.[37] In other words, we need to look at how organisations work with and implement AI, and how they decide exactly where and how to implement technology; we need to look at internal systems and controls; and we need to work on broadening public understanding of the technology so people better understand its limitations and benefits, and can get beyond the hype. Our approach to discussing AI risks and safety, and developing regulation and governance, should proceed with this richer conceptualisation of how problems with AI happen, and what can be done to reduce risk.[38]

Emerging international approaches to mitigating AI risks

International approaches to mitigating AI risks are developing apace. Importantly, those approaches are **not confined to regulation**. Regulation is important (and we come back to it below), but we think it important to emphasise *other* elements of international approaches to mitigating AI risks that have received less policymaker attention:

1. Fostering research in AI ethics[39] and safety[40] research, including enabling national coordination and international collaboration on research;
2. General action to increase understanding of, and preparation for the technology across society, including via education and workforce preparation;
3. Additional steps to manage the use of AI in government; and
4. Risk-based regulation.

Research

One element is **research**. Intense global technical competition in the development of AI and its capabilities is not necessarily being matched by investments in research in the collection of issues known as 'AI Safety'.[41] Australia should be investing in and globally collaborating in research into risks and risk mitigation, including what is often referred to as 'AI ethics and safety research'.[42] This must include technical research, but also (1) socio-technical and interdisciplinary research into understanding risks and their sources,[43] (2) technical and non-technical methods for reducing and mitigating risks[44] and (3) measuring the success of those mitigating efforts.[45]

Significant research is needed, because there are many new questions. For example, most of the work on transparency in AI has been largely done mainly on Explainable AI (XAI) and interpretability for understanding the world, or specifically, abstracting and representing the models of the world (or specific tasks, situations, or phenomena to be understood). For example, a recent work by an ADM+S researcher and collaborators looked at the attributions of model features and parameters in the prediction outcomes[46] e.g. to explain the influence of salient image features and the model parameters in the prediction outcomes of image recognition tasks. This is referred by Yann LeCun as 'system 1' thinking,[47] borrowing ideas from Daniel Kahneman's concept of System 1 and System 2 thinking. However, Large Language Models (LLMs) and Multimodal Foundation Models (MFMs) have shown emerging capabilities for system 2 thinking, which includes reasoning, adapting, and planning, given the advancement of reinforcement learning techniques, and access to human feedback at scale. This is the core of decision-making. There is very little work on transparency and explainability in reinforcement learning and planning agents. Thus, understanding the operations of LLM and MFM agents requires abstractions and transparency of both system 1 and system 2 thinking, especially in the context of Meta AI's chief scientist's bold vision of autonomous intelligence.[48]

Research efforts in this space are unavoidably global. ADM+S notes that investment in AI Safety Research is an important part of the recent US announcements;[49] and the UK too has established an AI Safety Institute, which recently released e.g. an AI safety evaluation platform.[50] We note the US-UK Memorandum of Understanding dated 1 April 2024, to develop tests for the most advanced AI models, following commitments from the AI Safety Summit. Australia should seek to engage with such efforts.[51] Australia should not only invest directly, but take any actions necessary to ensure that Australian researchers can collaborate with their global counterparts. This includes collaboration under large international grant schemes, facilitated by aligning our research funding schemes with equivalents overseas to facilitate joint investment.[52]

Efforts to prepare people, systems and society

A second element of international efforts is around socio-technical preparedness. Technology is embedded in, and affected by, its social and economic context. Emerging international approaches to both taking advantage of the potential of AI *and* mitigating risks involve a range of social efforts: for example, education and capacity uplift and workforce preparation. International efforts to address AI risks are also recognising the potential of AI to increase inequality across society. Australian efforts must do the same, including in particular by addressing digital exclusion.[53]

Government use of AI

A **third** element of international approaches to mitigating AI risk is addressing **government use of AI,** through a range of legal and extra-legal approaches**.** International approaches to the regulation of AI across the EU,[54] US, and Canada indicate an emerging consensus that government use of AI, especially in determining access to government services, is a high-risk use of AI that should be regulated. Canada for example has prioritised regulation of

government use of AI with its *Directive on Automated Decision-Making*,[55] which mandates a set of processes that government must adopt in developing and deploying AI.

The US Federal Government is also ensuring that government AI systems go through an assessment of risk, and that there is a record of systems deployed and their function. However, to ensure that agencies also see AI as a tool for modernizing operations and improving service, the Executive Order of 30 Oct 2023[56] and the Memo of the Office of Management and Budget of 28 March 2024 that expands on that Executive Order[57] have specific mandates to increase the agencies' capacity to responsibly adopt AI, above all requiring all agencies to designate a Chief AI Officer (CAIO) by 28 May 2024. Notably, the EO and Memo apply retrospectively to all systems already in place, and systems must be discontinued if they do not follow the compliance requirements and are not listed in the inventory.

Developing Regulatory Approaches

A fourth element is making adjustments to the overall regulatory environment, including through **risk-based regulation** with both *ex ante* obligations (such as risk assessment and mitigation) and ongoing elements (such as monitoring). There is a consensus around the appropriateness of *ex ante,* risk-based regulation. Other countries are also reviewing and updating existing laws, and ensuring that regulators have the resources, and expertise, to monitor developments and to act.

In relation to risk-based regulation, a number of other countries have already taken action. These efforts have features in common. The European, Canadian and US approach to government AI systems also align on baseline responsible practices for high-risk AI systems including documentation, logging, risk assessment and risk management, requirements of explanation, data governance, validation, testing, and ongoing monitoring. However, the systems introduced vary in their details. Australia is in a position to learn from the different approaches and take the best from the range of systems already in place.

Canada's approach as exemplified in its *Directive on Automated Decision-Making* is process-based regulation, rather than product-based regulation. Generally, these processes make it more likely that AI systems will be fairer, more transparent, and that there is more accountability around automated decisions. The Directive is technology neutral, being mainly concerned with the automation aspect of decision-making regardless of the technology used (AI or other forms of automation). Because the Directive is simple and clear, it is an appealing blueprint for a first step toward AI regulation in Australia. However, one potential issue with the Directive is that it tends not to focus on outcomes. There are few mechanisms that directly prohibit particular kinds of harmful outcome, or particular harmful uses of AI.

By contrast, Europe's AI Act prohibits applications of AI that are deemed unacceptably risky or harmful, including overbroad social-scoring, manipulation, emotion recognition, untargeted scraping of biometric information, and certain kinds of real time face recognition. It also establishes requirements for high-risk AI applications that go beyond processes, and address outcomes. The risk management requirement for high-risk systems is not just a box-

checking exercise, operators are required to reduce risks to 'acceptable' levels before high-risk systems are put into use. The disadvantage with the EU AI Act is primarily in its very complex structure and drafting, which is a product of the particular institutional setting.

Europe also relies on **standards** written by standards bodies to provide guidance on the implementation of regulatory requirements. One problem with this approach (especially for Australia, that has not relied significantly on standards in governance historically) is that the questions that arise in AI regulation bring up very difficult questions of human rights and the public interest. For example, when are risks from an AI system used in law enforcement 'acceptable'?[58] What kind of explanation should accompany an automated decision about access to an essential service such as insurance? While standards may be very effective in establishing responsible approaches to technical aspects of AI systems, it is not clear that standards bodies have the legitimacy or expertise to offer guidance on these difficult socio-technical questions,[59] especially if we want AI regulation to go beyond *process* to address and mitigate harmful *outcomes.*

International interoperability is important: Australian moves to regulate AI will need to coordinate, or be aware of, and even take advantage or join in regulatory moves overseas. We recommend strong participation in international cooperative mechanisms to manage risks arising from the largest models and actors, whether at a treaty level or technical standard-setting, to address issues regarding larger models at a global level. Australia already participates in technical standard-setting efforts and global discussions regarding the development of common principles and cooperation in their implementation. There is an opportunity to work with governments in a similar position - across countries like Canada, New Zealand, Singapore, Japan - with a strong interest in ensuring an open digital economy but coupled with genuine protection for competition and individuals. Australia's access to a sophisticated workforce and expertise, and large trade relationship with China, values aligned with the EU and Canada, good relationships with the US and UK and regional relationships in the Pacific means it is well placed to play an important role in such discussions.[60]

Opportunities to beneficially adopt AI, in ways that benefit citizens, the environment and/or economic growth, for example in health and climate management

There are a multitude of opportunities for AI to do good. One area worth highlighting, based on ADM+S' own research, is that Australia's for-purpose, social enterprise, social service and not-for-profit organisations are well-placed to lead the responsible and ethical adoption of AI. The sector's frontline role in addressing social disadvantage and inequality and providing essential advocacy for marginalised and at-risk groups, positions it as vital in ensuring that the benefits of AI flow through to all Australians. AI tools help organisations to improve their ability to monitor and understand social disadvantage or climate risks, and will likely improve efficiencies in service provision. While all of the risks addressed in this document are heightened for disadvantaged populations, the not-for-profit sector can play an important role in ensuring beneficial and inclusive AI adoption.

The not-for-profit sector and social purpose sector has mirrored the private and public sector in transforming operations and services through digital platforms and improved data collection and use.[61] AI and automation can build on these transformations to help organisations do more with less. Organisations are looking for accessible AI tools to help process unstructured qualitative data to improve services and outcomes. They are exploring applications for supporting research, and fast tracking grant application work where these involve repetitive or criteria-based processes.

In Australia's humanitarian sector, AI tools are helping to inform decision making in disaster preparedness, response and recovery. ADM+S researchers have been working with the Australian Red Cross to develop models for designing and deploying data-driven and ADM systems with vulnerable communities to improve resilience and preparedness.[62] The Centre has also worked with the Red Cross on questions around the use of AI and automation in aid and humanitarian situations with a focus on the role that facial recognition may play in identification processes.[63]

ADM+S researchers are also pursing 'AI for good' in other contexts. For example, we have researchers collaborating with international research institutes including the Centre for Human Compatible AI at UC Berkeley, and OpenAI, to undertake research looking at the ways Generative AI models might be used to inform more ethically nuanced automated decision making in contexts such as social media feed recommendations[64] and delegation of human responsibilities to AI agents.

Opportunities to foster a responsible AI industry in Australia

Throughout this submission we have highlighted significant developments in AI. All these developments present tremendous opportunities for Australia to build Australian-centric AI capabilities of human behaviour understanding and AI alignment, *using AI*, in conjunction with building methods and techniques for safe and responsible AI. *Responsible* innovation, consistent with Australian interests and values is critical. Australia has an outstanding group of world-leading researchers in this area, yet investment has been very scarce and simply not in the same ballpark with even other countries of comparable size and/or wealth. Australia needs to focus on building our own sovereign AI capabilities, collaborating with partner countries, rather than relying on imports of AI, and the global tech industry. The risks to Australia and Australians are exacerbated if we stop building in Australia, and become reliant on these technologies imported into our shores, without full understanding of these capabilities, and without building our own AI and deploying it responsibly in partnership with government agencies, educators, industry, and NGOs.[65]

Environmental impacts of AI technologies and opportunities for limiting and mitigating impacts

Environmental Impacts

Key environmental impacts of AI are diverse and distributed across their whole lifecycle from the development and training of AI models to their inference and application in a multiplicity of use cases.[66] There is growing concern about the **high energy use** - and hence carbon emissions - of the computing power needed not only to train and refine AI models, but also to store and transfer data, and to run AI applications in software and devices in businesses, homes and personal devices.[67] It has been estimated that the current relatively modest share of global carbon emissions represented by the information and communications sector (1.4%)[68] will blow out by 2040 to 14% or more, due in large part to the growth of data centres and digital technology including AI.[69] As this submission was being finalised, Microsoft published its latest sustainability report, showing a 30% *increase* in carbon emissions since 2020, due largely to its data centres.[70]

Use of water to cool the computing centres in which data is stored and models trained and in energy generation is also a major impact of increasing concern.[71] For example research published in the OECD AI Policy Observatory has estimated that, in Australia, every 26 queries of GPT-3 would evaporate approximately 500ml of water.[72] Both energy and water use are likely to become topics of contestation with AI services competing with other business and civic uses in particular locales for renewable energy and sometimes scarce water resources.[73]

Other major impacts in the AI lifecycle include raw material use – including the extraction and refining of rare metals to be used in the production of necessary equipment like graphics processing units;[74] land use for the siting of facilities like mines, processing and manufacturing plants for resources and hardware and data centres for compute;[75] and undersea cables (for data transfer).[76] The underlying logic of AI uptake also almost necessarily involves increasing proliferation and regular upgrading and replacement (and hence production and waste) of hardware from graphics processing units in data centres to the proliferation of business, home and personal mobile devices that incorporate AI applications.[77] The e-waste created by the regular upgrading of equipment to keep up with expanding AI development and applications will be an increasing contributor to Australia's existing waste and recycling challenge.

The impacts are similar *in kind* to those created by existing practices built on computing and data centres (e.g. cloud storage and digital platform operations).[78] However, the size and complexity of AI models developed, and the frequency of use and speed of development are likely to significantly heighten the demand for and use of energy and water for computing for the purposes of AI.[79]

The consideration of these impacts will become more critical with increasing global expectations that all public and private sector organisations fully report their carbon impact

throughout their whole supply chains; more acute pressure over the extraction and supply of silicon and critical minerals; and growing expectations that Australia move to a fully circular green economy. [80] In particular, as international law, international trade agreements and general market expectations all inexorably move towards more onerous carbon reporting regimes, AI providers will be expected to be able to provide information and logging capabilities about the carbon emissions associated with the development and running of the AI systems that will be embedded in software applications throughout the economy. Other environmental impacts, particularly biodiversity impacts (e.g. of mining, manufacturing and e-waste), are also likely to receive increasing attention in coming years as reporting frameworks for nature disclosures, biodiversity offsets, and circular economies come online.[81]

Opportunities for limiting and mitigating impacts

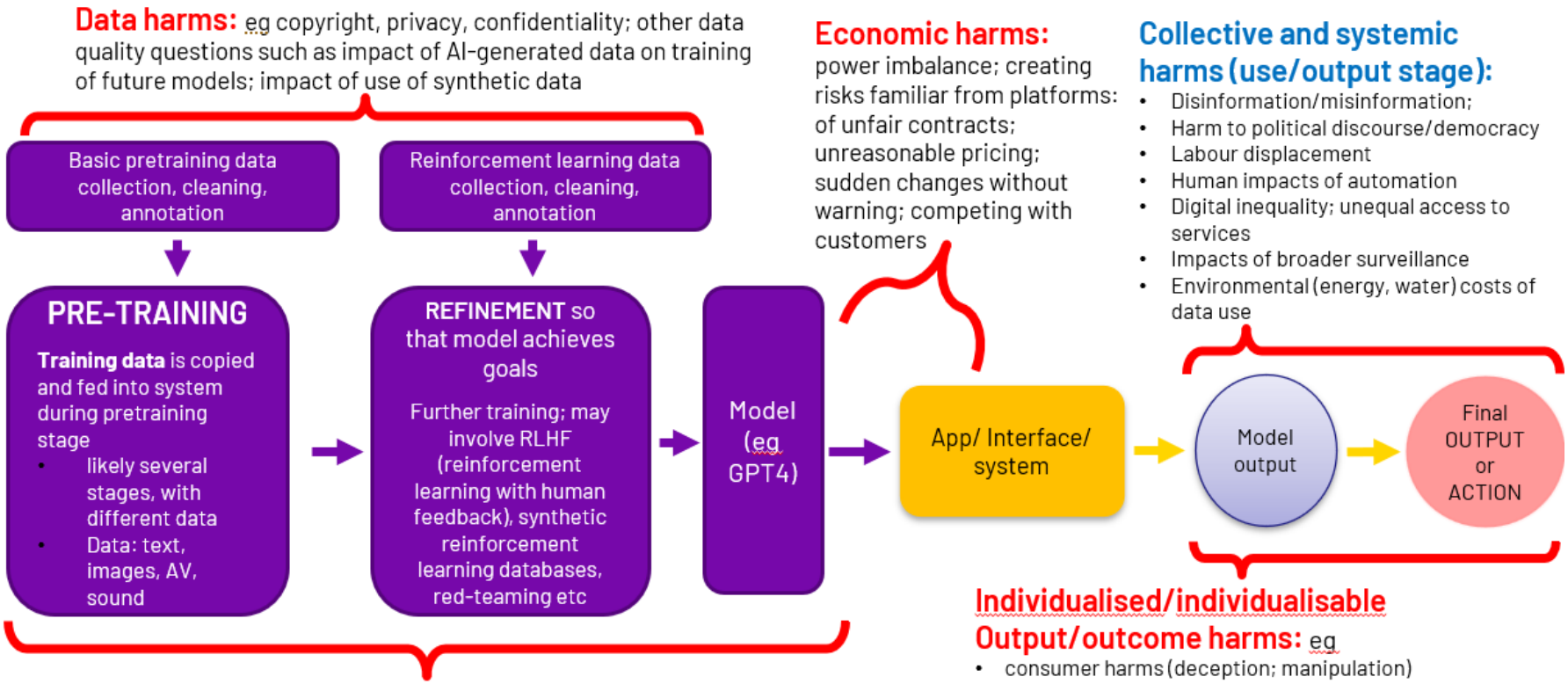
In light of these environmental impacts, environmental impact and sustainability should be defined as included within the broad concept of AI safety for the purposes of risk assessment and mitigation. [82] For example, the OECD's principles for responsible AI include recognition that both positive and negative impacts of AI on the environment need to be considered.[83] Australia could follow the approach of the EU AI Act which embeds environmental protection as a defined element of the overall goals of 'ensuring a high level of protection of healthy, safety [and] fundamental rights' in the uptake of AI (Article 1 and Annex IV).[84] The Act encourages the creation and implementation of voluntary codes of conduct for assessing and minimising environmental impact for all AI developers and providers (Articles 95(2), 112(7)). High-risk AI systems are expected to conduct risk assessments and create technical documentation processes that address their impact on 'health, safety [and] fundamental rights' (Articles 9, 11, and Annex IV), which (as noted above) is defined to include environmental protection . This includes a concern with the potential impact of A on a safe environment where AI is utilized in managing major facilities and infrastructures such as dams, electricity and manufacturing facilities that could have major environmental impact (by fire, flood, or pollution) if the AI fails and causes a major catastrophe. It also includes the environmental impact of the AI itself.

There is also a need to **review and update existing Australian laws and policies** to ensure they are fit for the purpose of ensuring environmentally responsible and sustainable AI, as AI applications are taken up across the whole of the public and private sector. Different legal and policy frameworks will touch on AI's environmental impact across different parts of the whole AI lifecycle including: environmental planning laws and licensing regimes for the siting and running of facilities such as mines and processing facilities for critical minerals, and data centres and undersea cables for data storage and compute power; energy grids including the creation and use of renewable energy facilities; carbon reporting and ESG frameworks; policies to incentivise and obligate product stewardship and e-waste reduction and re-use.

The relevance of environmental impacts extends also to soft law instruments. There are strong moves towards the development of technical standards for AI environmental impact logging and reporting in Europe and the US, and Australia should encourage and promote if not mandate the development of environmental impact logging and transparency standards

to support environmental reporting and transparency across the whole AI supply chain - since as public and private actors embed AI applications in their operations and businesses, they will need to be able to account for the significant carbon and other environmental impacts of the whole lifecycle of the AI and the hardware it is based on in their own environmental reporting. To support such standards, multiple groups in academia, advocacy, software engineering and industry are working on the quantification of carbon impacts of AI in application software,[85] as well as more holistic assessments of the environment impacts of AI systems.[86] In the US the proposed Artificial Intelligence Environmental Impacts Act of 2024 would 'direct the [US] National Institute of Standards and Technology (NIST) to develop standards to measure and report the full range of artificial intelligence's (AI) environmental impacts' including energy consumption and pollution across the full AI lifecycle.[87] The standard would form a basis for voluntary, and potentially mandatory, AI environmental auditing and reporting. An earlier version of the EU AI Act included similar provisions and it is still implicit in the approach adopted by the EU AI Act.[88]

Appendix A. Diagram of AI harms



Appendix B. ADM+S research outputs relevant to the Committee's Terms of Reference

ADM+S research addresses many of the questions raised by the Committee's Terms of Reference. Below is an incomplete listing of research papers and outputs that may be of relevance to the Committee, with abstracts/summaries to assist the committee in identifying any that are of interest. We are happy to put the researchers in touch with the committee or provide copies of any of the below on request.

Contact	Research	Description
T.J. Thomson ADM+S Affiliate (t.j.thomson@rmit.edu.au)	Australian media need generative AI policies to help navigate misinformation and disinformation	New research into generative AI images shows only over a third of media organisations surveyed at the time of research have an image-specific AI policy in place
	Generative visual AI in newsrooms	This essay explores key considerations that journalists and news organizations should be aware of when conceiving, sourcing, presenting, or seeking to fact-check AI-generated images
Georgia van Toorn ADMS Assc. Inv. (g.vantoorn@unsw.edu.au)	United against algorithms: a primer on disability-led struggles against algorithmic injustice	This primer explores how people with disability are collectively responding to the threats posed by algorithmic, data-driven systems - specifically their publzic sector applications.
Aaron Snoswell ADM+S Research Affiliate (a.snoswell@qut.edu.au)	Sociotechnical Specification for the Broader Impacts of Autonomous Vehicles	Lays out axes for broadening research in Autonomous Vehicle safety by incorporating more diverse socio-technical considerations.

	Reward Reports for Reinforcement Learning.	We propose a new AI procurement and documentation framework for mapping stakeholders and ongoing monitoring of emerging and dynamic threats from AI systems where user feedback is a key component.
	Draft G7 Guiding principles for organizations developing advanced AI systems	The Queensland University of Technology Centre for Data Science has prepared this submission to the draft Guiding Principles for Organizations Developing Advanced AI, prepared jointly by the EU and the G7 nations.
Aaron Snoswell Jean Burgess; Nicolas Suzor	Measuring Misogyny in Natural Language Generation: Preliminary Results from a Case Study on two Reddit Communities	Studies the problem of measuring toxicity generally and misogyny specifically in natural language generation, highlighting severe shortcomings in contemporary technical approaches.
Henry Fraser ADM+S Research Affiliate Aaron Snoswell; Rhlye Simcock	AI Opacity and Explainability in Tort Litigation	We propose a novel regulatory approach to mitigating AI opacity in Negligence lawsuits, combining technical and socio-technical approaches. This framework has in part been adopted in recent EU product liability legislation.
Henry Fraser	'AI Safety' doesn't make AI safe	Research into the more-than-technical dimensions of AI harms, and the inadequacy of an 'error-based' concept of AI risks and harms.
Ariadna Matamoros-Fernández (ariadna.matamorosfernandez@qut.edu.au)	Mapping and critiquing the new vendor ecosystem in the 'Trust and Safety' Industry	Research into the emerging vendor ecosystem offering AI as a service, especially those that offer AI as a content moderation solution.
Thomas Graham (ADMS Researcher) thomasgraham@swin.edu.au	Navigating AI-lien Terrain: Legal liability for artificial intelligence in outer space	The paper investigates the application of UN space treaties, regional AI regulations, and industry initiatives to space-based AI systems.

Julia Stoyanovich ADMS Partner Inv. (stoyanovich@nyu.edu)	ADM+S Partner Investigator invited to speak about responsible AI at the United Nations	ADM+S Partner Investigator Prof Julia Stoyanovich from New York University was invited to speak on responsible artificial intelligence (AI) at the United Nation's 62nd session of the Commission for Social Development (CSocD62).
Dr Monika Zalnieriute Associate Investigator, (m.zalnieriute@unsw.edu.au)	Glukhin v. Russia	Glukhin v Russia is the first ECtHR decision on the use of FRT; it portends a strong foundation for further restricting how governments use FRT.
Anthony McCosker ADM+S CI (amccosker@swin.edu.au)[3rd Author]	Using Data Bricolage and Mixed-Methods GIS to Uncover Mental Health Service Gaps in Rural Australia	Reapplying the methods within this work within an AI or NLP process can enable much larger data sets to be analyzed in real time, offering greater insight for NFP decision makers.
	Towards resilient communities: data capability and resource mapping for disaster preparedness	In this project, the research team based in the ADM+S worked with the Australian Red Cross to better understand the challenges and potential of data-driven decision-making for community disaster resilience.
	Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube	Examining the rise and threat of deepfakes (non-consensual synthetic media content), this paper moves beyond the three traditional types of response - regulatory, technical and media literacy - to model data and AI literacy as important forms of social responses.
Chris O'Neill ADM+S Research Fellow (chris.oneill@monash.edu)	Disaster, facial recognition technology, and the problem of the corpse	Drawing upon interviews conducted with senior Australian government administrators and researchers, I argue that FRTs are being used to respond to the trauma of disaster through its novel mediation and refiguration, tied to discourses of resilience which have been used to justify the expansion of FRT as a means for relief and the provision of aid

PRO (YONCHANOK) KHAOKAEW ADMS Researcher (y.khaokaew@unsw.edu.au)	MAPLE: Mobile App Prediction Leveraging Large Language Model Embeddings	This study introduces a novel prediction model, Mobile App Prediction Leveraging Large Language Model Embeddings (MAPLE), which employs Large Language Models (LLMs) and installed app similarity to overcome challenges in AI app recommendation, user behaviour analysis, and mobile resource management
	ZzzGPT: An Interactive GPT Approach to Enhance Sleep Quality	This paper delves into the role of technology in understanding sleep patterns. We introduce a two-stage framework, utilizing Large Language Models (LLMs), aiming to provide accurate sleep predictions with actionable feedback.
Yong-Bin Kang is a Research Fellow (ykang@swin.edu.au)	AIoT-CitySense: AI and IoT-Driven City-Scale Sensing for Roadside Infrastructure Maintenance	This paper presents AIoT-CitySense, an AI and IoT-driven city-scale sensing framework, developed and piloted in collaboration with a local government in Australia
Sarah Pink ADMS CI (sarah.pink@monash.edu)	Trust, artificial intelligence and software practitioners: an interdisciplinary agenda	In this article we develop an interdisciplinary approach, using socio-technical software engineering and design anthropological approaches, to investigate how trust and trustworthiness concepts are articulated and performed by AI software practitioners.
	Design anthropological foresighting: Reframing automated futures	Applies a new design and futures anthropology theory to qualitative foresighting.
Christopher Leckie ADMS CI (caleckie@unimelb.edu.au) [3rd Author]	Adversarial Coreset Selection for Efficient Robust Training	Developing an approach to speed up adversarial training by 2–3 times while experiencing a slight degradation in the clean and robust accuracy.
Jake Goldenfein ADM+S CI	Privacy's Loose Grip on Facial Recognition	Book chapter discussing privacy's inability to manage dataset economies in the facial recognition context.

jakeg@unimelb.edu.au	Lost in the Loop: Who is the "Human" of the Human in the Loop	Book chapter outlining how legal requirements for a human in the loop produce troubling distributions of accountability and responsibility in government contexts.
	GenAI Concepts	Web resource of the public and public service, outlining some of the technical, operational, and regulatory concepts and complexities around the uptake of AI by governments.
Simon Coghlan ADMS Affiliate (simon.coghlan@unimelb.edu.au) Christine Parker ADMS CI (christine.parker@unimelb.edu.au)	Helping and not Harming Animals with AI	Ethical discussions about Artificial Intelligence (AI) often overlook its potentially large impact on nonhuman animals…we argue that there is some value in focusing on technology-based harms in the context of AI ethics and policy discourses
Melissa Gregg ADM+S Board Member & Senior Industry Fellow melissa.gregg@rmit.edu.au Yolande Strengers ADMS Assoc. Inv. (yolande.strengers@monash.edu)	Getting beyond Net Zero dashboards in the information technology sector	Drawing on first-hand observation of this practice over several years of employment in a US multinational firm, we focus on the carbon emissions dashboard as a representative example of key problems with the dominant approach to sustainability targets in the Information Technology (IT) sector.

Endnotes

[1] Molly McHugh-Johnson, '100 things we announced at I/O 2024', *The Keyword* (Blog, May 2024).

[2] 'Hello GPT-4o', *OpenAI* (Web Page, May 2024).

[3] This would appear to be another example of generalised multimodal AI models surpassing specialised AI models in performance, something we also saw with Google's Gemini.

[4] 'Meet Your New Assistant: Meta AI, Built With Llama 3' , *Meta* (Blog, April 2024).

[5] Tom Warren, 'Microsoft 365's Copilot gets a GPT-4 Turbo upgrade and improved image generation', *The Verge* (Web Page, April 2024).

[6] 'Project Astra', *Google* (Web Page, May 2024).

[7] Dang Nguyen et al. "AI and Automated Decision-Making in News and Media." Report. ARC Centre of Excellence for Automated Decision-Making and Society, December 5, 2023. Australia. . See also Matamoros-Fernández, A et al., (2021). What's "Up Next"? Investigating Algorithmic Recommendations on YouTube Across Issues and Over Time. Media and Communication, 9(4), Article 4.

[8] Angus, D et al. (2024) The Australian Ad Observatory Technical and Data Report, *ADM+S Working Paper Series 2024* (9), ARC Centre of Excellence for Automated Decision-Making and Society.

[9] See *Australian Internet Observatory* and *Australian Research Data Commons*.

[10] Academy of the Social Sciences in Australia. (2024) *Decadal Plan for Social Science Research Infrastructure 2024-33*.

[11] Salar Abbaspourazad et al., (2024) 'Large-scale Training of Foundation Models for Wearable Biosignals', Machine Learning Apple.

[12] See OpenAI, 'Introducing GPT 4o', YouTube Video (see at 23:38).

[13] See Thomas, J., McCosker et al, ARC Centre of Excellence for Automated Decision-Making and Society, RMIT University, Swinburne University of Technology, and Telstra, *Measuring Australia's Digital Divide: Australian Digital Inclusion Index: 2023* (Report, 2023); and Featherstone D, Ormond-Parker L, Ganley L, Thomas J, Parkinson, S, Hegarty K, Kennedy J, Holcombe-James I, Valenta L, Hawkins, L (2023) *Mapping the Digital Gap: 2023 Outcomes Report*, Melbourne: ARC Centre of Excellence for Automated Decision-Making and Society.

[14] Madhumita Murgia and Richard Waters, 'Google Considers Charging for AI-powered search in big change to business model', *Financial Times*, 4 April 2024.

[15] Weatherall, Kimberlee, Paul Henman, Jose-Miguel Bello y Villarino, Rita Matulionyte, Lyndal Sleep, and Melanie Trezise. "Automated Decision-Making in New South Wales: Mapping and Analysis of the Use of ADM Systems by State and Local Governments." Report. ARC Centre of Excellence for Automated Decision-Making and Society, March 8, 2024. New South Wales.

[16] NSW Ombudsman, A map of automated decision-making in the NSW Public Sector: A special report to Parliament (2024).

[17] Researchers at ADM+S have worked with the Office of the Victorian Information Commissioner to produce GenAI literacy resources for public sector stakeholders: see Fan Yang, Jake Goldenfein, Kathy Nickels, 'GenAI Concepts'.

[18] Weatherall et al, ADM+S submission to the Safe and responsible AI in Australia discussion paper (ADM+S, 2023).

[19] Bartolo, L., & Matamoros-Fernández, A. (2023). Online Harm. ISP-Platform Governance Terminologies Essay Series.

[20] See the OECD's Revised definition of AI: 'An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment': Explanatory Memorandum on the Updated OECD Definition of an AI System. OECD Artificial Intelligence Papers. Vol. 8. OECD Artificial Intelligence Papers, March 5, 2024.

[21] Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2016) 3(1) *Big Data & Society*. 'Explainability' is an ongoing area of research: far from a solved problem in relation to AI and becoming more challenging with newer generations of technology.

[22] US National Institute of Standards and Technology (NIST), *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (April 2024) defines generative AI as 'the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text and other digital content' (p1, fn1).

[23] 'Information with integrity can be trusted, distinguishes fact from fiction, opinion, and inference, acknowledges uncertainty and is transparent about its level of vetting': NIST.

[24] Anthropic, 'Measuring the persuasiveness of language models', 10 April 2024.

[25] McCosker, A. (2024). Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. *new media & society*, *26*(5), 2786-2803.

[26] For example, there is a largely unregulated burgeoning ecosystem of vendors that are building tools off pre-existing multimodal foundation models (e.g. GPT-4, Gemini, etc) and offer them as content moderation solutions capable of identifying and taking down harmful content at scale; systems offering automated content moderation solutions for clients seeking to comply with legislation but without in-house trust and safety teams: Matamoros-Fernández, A. & Bartolo, L. (2024). Mapping and critiquing the new vendor ecosystem in the 'Trust and Safety' Industry. Selected Papers in Internet Research 2024. Association of Internet Researchers.

[27] Jake Goldenfein, 'Privacy's Loose Grip on Facial Recognition: Law and the Operational Image' in Rita Matutlionyte and Monika Zalnieriute (eds) The Cambridge Handbook of Facial Recognition in the Modern State (Cambridge 2024); Kimberlee Weatherall, 'IP and data, IP in data, IP as data', in Damian Clifford, Jeannie Paterson and Kwan Ho Lau (eds) Data Rights and Private Law (Hart Publishing 2023).

[28] Evan Hubinger et al, 'Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training' (No arXiv:2401.05566, arXiv, 17 January 2024) <http://arxiv.org/abs/2401.05566>

[29] See the table in the ADM+S submission to the Safe and responsible AI in Australia discussion paper, at 13-15.

[30] Notably, reforms to Australian consumer protection law (in the form of a prohibition on unfair trading practices proposed by the ACCC); stronger privacy protections (proposed in the *Privacy Act Review*, such as rights fair and reasonable data processing); rights to reasons and information about automated decision-making (proposed by both the Australian Human Rights Commission *Human Rights and Technology* Report and the Royal Commission on Robodebt.

[31] More gaps in Australia's capacity for enforcement - for example, the barriers to direct enforcement of both human rights protections and privacy protections - are summarised in our ADM+S submission to the Safe and responsible AI in Australia discussion paper, 17-18.

[32] 'Henry Fraser and Nicolas P Suzor, Locating Fault for AI Harms: A systems theory of foreseeability, reasonable care and causal responsibility in the AI value chain' (Working Paper, April 2024).

[33] Department of Industry, Science and Resources, *Australia's Government's Interim Response to Safe and Responsible AI in Australia Consultation* (Interim Report, 2024).

[34] Above n 19.

[35] Bobby Allyn, 'Google races to find a solution after AI generator Gemini misses the mark' NPR.

[36] Henry Fraser, '"AI Safety" Doesn't Make AI Safe' (The Learning Curve, 28 June 2023) accessed 11 March 2024.

[37] Roel IJ Dobbe, 'System Safety and Artificial Intelligence' in Justin B Bullock and others (eds), *The Oxford Handbook of AI Governance* (Oxford University Press 2022).

[38] Henry Fraser and Nicolas P. Suzor, 'Locating liability for AI harms: A systems theory of reasonable foreseeability, control and fault in the AI value chain' (Working Paper, April 2024).

[39] FAccT, Statement on AI Harms and Policy (Statement).

[40] Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute.

[41] Emerging Technology Observatory, *The state of global AI safety research* (Web Page, 2024); Stanford University Human-Centred Artificial Intelligence, *Artificial Intelligence Index Report 2024* (Report).

[42] Ibid.

[43] See, e.g. Dean, Sarah, et al. "Axes for sociotechnical inquiry in AI research." IEEE Transactions on Technology and Society 2.2 (2021): 62-70. On the technical side, this could include, for example, investment in research into understanding how new capabilities emerge in larger models, for better anticipation of emerging risks. More broadly, it could include research, as mentioned earlier in this submission, into how humans interact with newer models, frameworks for mapping stakeholders, risks, and opportunities in the deployment of new models, and whether newer forms of interaction increase risks of manipulation or vulnerability.

[44] See, e.g. Gilbert, Thomas Krendl, et al. "Sociotechnical Specification for the Broader Impacts of Autonomous Vehicles." arXiv preprint arXiv:2205.07395 (2022). Gilbert, Thomas Krendl, et al. "Reward reports for reinforcement learning." Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. 2023. This could include, for example, investment in methods of model alignment, and methods for protecting models from a range of attacks, such as prompt injection attacks or attacks that seek to extract underlying data, methods for watermarking and/or detection of AI generated content, authenticity and/or copyright verification methods, etc.

[45] This could include, for example, research into methods for detecting capabilities that create risk, such as whether a model may take action to achieve a goal or reward by deceiving a human user.

[46] Zhiyu Zhu et al., AttEXplore: Attribution for Explanation with model parameters eXploration, International Conference on Learning Representations (ICLR), 2024.

[47] Yann LeCun, A Path Towards Autonomous Machine Intelligence (2022).

[48] 'Yann LeCun on a vision to make AI systems learn and reason like animals and humans', *Meta* (Blog, February 2022).

[49] US Senate, *Driving US Innovation in Artificial Intelligence* (Roadmap, 2024).

[50] UK GOV, 'AI Safety Institute releases new AI safety evaluations platform' (Media Release, May 2024).

[51] UK GOV, 'UK & United States announce partnership on science of AI safety' (Media Release, April 2024).

[52] We note, in this connection, the current review of the National Competitive Grants Schemes.

[53] Above n 13.

[54] In the case of the EU, it is notable that many of the uses specifically identified as high-risk involve public sector uses (for example, in law enforcement). See European Parliament, *European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD) ('EU AI Act')*. See the text.

[55] Government of Canada, Directive on Automated Decision Making (April 2023).

[56] The White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (October 2023).

[57] The Office of Management and Budget, Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence (October 2023).

[58] Henry Fraser and José-Miguel Bello y Villarino, 'Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough', *European Journal of Risk Regulation*, 2023, 1–16 http://dx.doi.org/10.1017/err.2023.57.

[59] Henry Fraser, Christine Parker, Fiona Haines, Kimberlee Weatherall and José-Miguel Bello y Villarino, 'What Role for Standards and Assurance in Regulating Artificial Intelligence in Australia?' (2023).

[60] José-Miguel Bello y Villarino,, David Hua, Barry Wang, Melanie Trezise, (2023) 'Standardisation, trust and democratic principles: the global race to regulate Artificial Intelligence', *United States Studies Centre*.

[61] McCosker A., Shaw F., Yao X., Albury K. (2022) A Data Capability Framework for the Not-for-Profit Sector. Swinburne University, Melbourne.

[62] McCosker, A., Kang, Y. B., Shaw, F. (2023) 'Towards Resilient Communities: Data Capability and Resource Mapping for Disaster Preparedness', Swinburne University of Technology and ARC Centre of Excellence for Automated Decision Making and Society, Melbourne.

[63] See eg O'Neill, C. (2024). 'Disaster, facial recognition technology, and the problem of the corpse.' *New Media & Society*, *26*(3), 1333-1348.

[64] Rebecca Ralph, GenAI Lab team makes the finals of the Prosocial Ranking Challenge, QUT (Blog, May 2024).

[65] Other liberal democracies such as France or Spain are already ensuring this onshoring with their own LLMs, either through investment (France in Mistral) or developing new public and open models with a technology partner (Spain with IBM). A Centre's researcher has suggested starting those developments with Multi Modal Models for education, using data already held by public entities. Bello Villarino, J.M (2023). An AI foundation model for education. In Tomaževiè, N., Ravšelj, D., Aristovnik, A. (Eds.), *Artificial Intelligence for human-centric society*, (pp. 114-133). Brussels: European Liberal Forum.

[66] Alexandra (Sasha) Luccioni, 'The mounting human and environmental costs of generative AI', *Ars Technica* (Op-ed, 12 April 2023). ; Ligozat, Anne-Laure, Julien Lefevre, Aurélie Bugeau & Jacques Combaz. 2022. 'Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions.' *Sustainability* 14, no. 9: 5172. https://doi.org/10.3390/su14095172; AlgorithmWatch, 'Digging Deeper: AI's Environmental Report Card. Does Artificial Intelligence Consume More Resources than it Conserves?' (2023) *SustAIn magazine, Issue #2*.

[67] Roel Dobbe and Meredith Whittaker, 'AI and Climate Change: How they're connected, and what we can do about it' (2019) *AI Now Institute*, Medium; Lynn H. Kaack et al, 'Aligning Artificial Intelligence with Climate Change Mitigation.'(2022) 12(6) *Nature Climate Change* 518–27; Roy Schwartz et al, 'Green AI.' (2020) 12 *Communications of the ACM 63* 54–63; Nafus, Dawn, Eve M. Schooler, and Karly Ann Burch. "Carbon-responsive computing: Changing the nexus between energy and computing." *Energies* 14.21 (2021): 6917.

[68] Jens Malmodin, Nina Lövehagen, Pernilla Bergmark, Dag Lundén, ICT sector electricity consumption and greenhouse gas emissions – 2020 outcome, Telecommunications Policy, Volume 48, Issue 3, 2024, 102701.

[69] Nafus, Dawn, Eve M. Schooler, and Karly Ann Burch. "Carbon-responsive computing: Changing the nexus between energy and computing." *Energies* 14.21 (2021): 6917.

[70] Microsoft, *2024 Sustainability Report*, (Report, 2024).

[71] Pengfei Li et al, 'Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models' (2023) arXiv: 2304.03271.

[72] Shaolei Ren, 'How much water does AI consume? The public deserves to know' OECD AI Policy Observatory (30 November 2023).

[73] Mel Hogan, 'Data flows and water woes: The Utah Data Center' (2015) 2(2) *Big Data & Society*.

[74] Ludovico Rella, 'Close to the metal: Towards a material political economy of the epistemology of computation' (2023) *Social Studies of Science*.

[75] Mel Hogan and A Vonderau, 'The nature of data centers' (2019) *Culture Machine*.

[76] Nicole Starosielski, *The Undersea Network* (Duke University Press, 2015).

[77] Gregg, Melissa, and Yolande Strengers. "Getting beyond Net Zero dashboards in the information technology sector." *Energy Research & Social Science* 108 (2024): 103397.

[78] Jesse Dodge et al, 'Measuring the Carbon Intensity of AI in Cloud Instances.' In 2022 *ACM Conference on Fairness, Accountability, and Transparency*, 1877–94.

[79] Alexandra Sasha Luccioni, 'The mounting human and environmental costs of generative AI', *Ars Technica* (Op-ed, 12 April 2023).

[80] See e.g. Microsoft, *2024 Sustainability Report*, (Report, 2024).

[81] For a useful discussion of how these various requirements are likely to impact environmental governance requirements for AI, see Phillip Hacker, Sustainable AI Regulation. *Common Market Law Review*, 61(2), 345–386 at 371–374.

[82] See e.g. Philipp Hacker, 'Sustainable AI Regulation' (2024) 61(2) *Common Market Law Review*; Rakova, Bogdana, and Roel Dobbe. "Algorithms as social-ecological-technological systems: an environmental justice lens on algorithmic audits." *arXiv preprint arXiv:2305.05733* (2023).

[83] Kaith Streier et al, 'Can AI help save the planet?' *OECD.AI Policy Observatory*, (Web Page, 17 November 2022); OECD, *Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint* (Report, 2022).

[84] European Parliament, European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)('EU AI Act'). Recital 1, Article 1. See also the text.

[85] E.g. Green Software Foundation; OECD Expert Group on Compute and Climate.

[86] Anne-Laure Ligozat, Julien Lefevre, Aurélie Bugeau, and Jacques Combaz, 'Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions' (2022) 14 (9) *Sustainability* 5172; Aimee van Wynsberghe, 'Sustainable AI: AI for Sustainability and the Sustainability of AI' (2021) 1(3) *AI and Ethics* 213–18; For an example see 'SustAIn: The Sustainability Index for Artificial Intelligence', *AlgorithmWatch* (Web Page).

[87] Senator Ed Markey, 'Markey, Heinrich, Eshoo, Beyer Introduce Legislation to Investigate, Measure Environmental Impacts of Artificial Intelligence,' Press Release, 1 February 2024.

[88] Above n 59 pp 15-16.