Supplementary Submission
No 1a

INQUIRY INTO ARTIFICIAL INTELLIGENCE (AI) IN NEW SOUTH WALES

Organisation:	Campaign for AI Safety
Date Received:	19 October 2023

Campaign for AI Safety

SUBMISSION TO THE PARLIAMENT OF NEW SOUTH WALES, THE PORTFOLIO COMMITTEE NO. 1 - PREMIER AND FINANCE

Inquiry into artificial
intelligence (AI) in
New South Wales

20 October 2023

Campaign for AI Safety

Campaign for AI Safety

Submission: Inquiry into Artificial Intelligence (AI) in NSW

This submission is made by the Campaign for AI Safety in response to the Portfolio Committee No. 1 - Premier and Finance inquiry into artificial intelligence (AI). We have responded to selected parts of the terms of reference.

The Campaign for AI Safety is a not-for-profit association established in Australia with members in Australia and other countries. We are concerned about the dangers AI poses to people and advocate for a stop on the advancement of certain AI capabilities. We also advocate for regulation that promotes and mandates ethical AI. We are not affiliated with any political group. Please visit campaignforaisafety.org for more information.

Campaign for AI Safety

Key points

- AI has improved people's lives overall and delivers social and economic benefits. We are deeply concerned that increasingly powerful and intelligent AI can be misused and/or cause large-scale societal harm. AI researchers and the industry share this concern and have warned us about these risks[1].

- The Commonwealth has responsibility for the majority of laws and regulations that regulate AI use. Our policy recommendations are focussed on areas of residual risk that NSW can act on to protect citizens from the most harmful effects of AI:

    - Strengthen consumers' right to redress and extending the product safety regime to AI technologies.

    - Restrict highly advanced AI and AI that exhibit signs of dangerous capabilities from being used by public entities, government service delivery and in critical infrastructure.

    - Guidance on safe and responsible use of generative AI by public servants.

    - Add safety as a principle in the NSW Government's AI Ethics Policy.

    - Redirect government funding in AI industry or state capability development to safety research.

    - Impose transparency and information disclosure requirements on the supply of high-risk AI systems in NSW.

    - If overseas jurisdictions such as the UK or USA licence AI developers for responsible development, restrict the supply of high-risk AI systems into NSW to developers licensed in those jurisdictions.

    - Prohibit the development of increasingly powerful and intelligent general-purpose AI in NSW.

[1] "Statement on AI Risk", Center for AI Safety (2023): signed by Sam Altman (CEO, OpenAI), Demis Hassabis (CEO, Google DeepMind), Kevin Scott (CTO, Microsoft) and others.

Campaign for AI Safety

1. AI risks to NSW community

We are most concerned about catastrophic risks from the use of large-scale general AI that exhibit signs of dangerous capabilities

The NSW community is already experiencing the negative impacts of AI such as its use to:

- violate people's privacy[2]
- create and spread disinformation (e.g. using videos of AEC vote counting staff to create narratives on social media about rigging of elections) [3]
- compromise cybersecurity (e.g. Latitude Financial breach of NSW driver licence information)[4]
- generate decisions that are biased (e.g. Robodebt automated decision making and risk assessments in criminal sentencing for minority groups in the US)[5]
- censorship (e.g. Meta's censorship of the Legalise Cannabis Party's logo[6])
- Cause physical and mental harms (e.g. school aged children in Australia using AI-generated deep fake pornography to bully peers[7])
- threaten livelihoods (AI is built on vast amounts of copyright information which threaten writers, artists and the creative industries' livelihoods[8] and has led to legal challenges overseas such as the GitHub Copilot litigation and the Getty litigation in the UK and US).

At the same time, there is increasing concern that highly advanced AI systems that exhibit dangerous capabilities could be misused to create societal-level catastrophic harms, of which we list a few:

- Hacking and cyberattacks

Attempts have been made to develop AI capable of hacking, such as WormGPT[9], a generative tool that can launch sophisticated phishing and business email compromise attacks. Cybersecurity researchers have demonstrated a variety of potentially malicious use

[2] "Clearview AI breached Australians' privacy", Office of the Australian Information Commissioner, OAIC and the UK's Information Commissioner's Office, ICO (3 November 2021).

[3] "'Stolen' federal election narratives saw TikTok asked by AEC to remove footage of vote-counting staff", Ariel Bogle, ABC News (Updated on 31 August 2022).

[4] "Latitude Financial breach", Service NSW, NSW Government (Updated on 13 September 2023).

[5] "The Flawed Algorithm at the Heart of Robodebt", Associate Professor Toby Murray, Dr Marc Cheong and Professor Jeannie Paterson, University of Melbourne (10 July 2023).

[6] "Artificial intelligence: First Australian parliamentary inquiry into AI to be led by pro-cannabis party", Nick Bonyhady, The Australian Financial Review (28 June 2023).

[7] "AI being used for child sex abuse images in regulation-free zone", Nick Bonyhady, The Australian Financial Review (15 August 2023).

[8] "Australian artists accuse popular AI imaging app of stealing content, call for stricter copyright laws", Cait Kelly, The Guardian (12 December 2022).

[9] "WormGPT: What to know about ChatGPT's malicious cousin", Charlie Osborne, ZDNET (20 July 2023).

cases. For example, a team at Claroty Ltd, a cybersecurity business, used ChatGPT to win a hacking tournament[10]. In March 2023, EuroPol (the law enforcement agency of the European Union) published explorations of how ChatGPT is able to facilitate a significant number of criminal activities, ranging from helping criminals to stay anonymous to specific crimes including terrorism and child sexual exploitation[11].

- Deception and social manipulation

The Alignment Research Center found that GPT-4 could pretend to be a blind person to hire a human via an online job ad to pass the CAPTCHA test so that it could access the internet[12]. Graphika, a research company that studies disinformation, has uncovered 'deepfake' video technology and AI-generated images in pro-China campaigns disseminated through social media to influence and manipulate people's views[13].

- Bioterrorism

A MIT experiment found that non-experts could use large language models (LLMs such as ChatGPT and Google's Bard) to identify, acquire and release viruses that could cause widespread harm. In one hour, MIT non-scientist students were given detailed instructions on how to engineer four potential pandemic pathogens and potential mistakes to avoid, and the chatbots named suppliers that were unlikely to verify orders[14].

The scientists who created Chemchrow, a GPT-4 powered tool that can execute common chemical tasks across areas such as drug and materials design and synthesis, acknowledge it can be repurposed for harmful applications, such as designing chemical weapons[15].

In 2022, researchers tweaked an existing biochemical research AI product[16] to reward toxicity: it produced molecules that could be deadlier than existing biochemical weapons.

Other examples are[17]:

- Weapons acquisition
- Long-horizon planning
- Situational awareness
- Persuasion and manipulation
- Self-proliferation.

[10] "ChatGPT Helped Win a Hackathon", Kim S. Nash, WSJ PRO (20 March 2023).
[11] "ChatGPT: The impact of Large Language Models on Law Enforcement", European Union Agency for Law Enforcement Cooperation, Tech Watch Flash (Updated on 17 April 2023).
[12] "Update on ARC's recent eval efforts", ARC Evals (Updated on 18 March 2023).
[13] "Deepfake It Till You Make It", GRAPHIKA (7 February 2023).
[14] "Chatbots allow people with no lab training to create pandemic viruses, study finds", Matthias Bastian, The Decoder by DEEP CONTENT (18 June 2023).
[15] "ChemCrow: Augmenting large-language models with chemistry tools", Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller (21 June 2023).
[16] "Dual use of artificial-intelligence-powered drug discovery", Fabio Urbina, Filippa Lentzos, Cédric Invernizzi and Sean Ekins (7 March 2022): Nature Machine Intelligence volume 4, pages 189–191.
[17] "Model evaluation for extreme risks", Toby Shevlane, et al. (24 May 2023).

In our view, the best approach to address all of the above risks (both the harms currently being experienced by NSW citizens and future, hypothetical catastrophic risks) is to regulate AI so that industry makes models safer and prevent development of more powerful and intelligent AI than currently exists. We make more detailed policy recommendations in the next chapter.

2. Recommendations to manage AI risks in NSW

The Commonwealth is currently considering whether further regulation of AI is required to protect Australians. In our submission to the consultation in August 2023, we called for:

1. existing laws and regulations to be updated to cover gaps and;
2. creation of new AI-specific laws and bodies to manage risks that we feel are very specific to AI and cannot be addressed by existing regulation.

We make detailed policy recommendations to cover the gaps we identified in consumer protection, copyright and government administration regulations that have the effect of regulating AI. We also propose a risk-based regulatory framework for the development of AI technologies and, separately, for the use of AI applications. We call for the Commonwealth to prohibit the use and development of highly advanced AI that are uninterpretable, agentic ("human-out-of-the loop" AI applications[18]) and exhibit signs of dangerous capabilities in Australia.

Our submission can be found on our website.

Commonwealth regulation of AI is probably the most effective approach to ensure appropriate safeguards are in place for NSW citizens as it is responsible for most of the laws that affect AI. Below, we identify policy levers the NSW Government can use to **mitigate any residual risks.**

Strengthen liability rules for harms caused with AI

The complexity of the AI supply chain and the opaque and autonomous nature of AI models' behaviour and decision making make it very challenging for individuals and small businesses to seek redress for damages or harms incurred under current liability and product safety laws.

If the rights of an NSW individual or customer are infringed, they should be able to sue not just the immediate party that employed the AI system (for example, a business using AI-powered recruiting software that has a bias against minorities), but also the provider of the AI system (e.g. the recruitment software provider or the API provider such as Amazon or Microsoft) as well as the AI lab that trained and released the AI system (e.g. Anthropic or OpenAI). This is consistent with other, 'traditional' industries in NSW where strict product safety and manufacturer liability rules allow for legal recourse for victims and guides innovation towards safer products.

Shine Lawyers[19] propose the following for law reform:

[18] "Human-in-the-loop", Wikipedia (Updated on 21 March 2023).

[19] See their submission for in-depth analysis and policy recommendations: "Shine's Submission to the Department of Industry, Science and Resources - Campaign for AI Safety", Atanaan Ilango and Dr Benjamin Koh, Shine's Class Actions practice (25 July 2023).

- In line with current EU proposals,[20] all AI developers should have a presumed duty to its end-users and non-contracting third parties for the harms their products have caused:
    - This is a rebuttable duty that applies only to significant harm.
    - This affords clarity to the general public (i.e. non-contracting third parties) on the legal recourse available to them in the event a software or AI product causes them harm.
    - The liability should include the injury of pure mental harm (e.g. embarrassment, stress) and pure economic loss.
- In line with recent changes to Australia's Unfair Contract Terms (UCT) laws in the *Competition and Consumer Act 2010* and *Australian Securities and Investments Commission Act 2001*, AI-related laws should specifically state that any terms within user-agreements for AI-related software that exclude liability or prevent an individual's right to participate in class actions are to be deemed UCT and voided.

The Commonwealth and states share responsibility for product safety regulation. There may be scope within NSW's fair trading laws to enact at least some or all these reforms. We refer to the EU's Product Liability Directive and AI Liability Directive that we feel are relevant and adaptable for NSW.

The benefit of making it easier for individuals to seek redress is that it will incentivise AI developers to prioritise safety and ensure accuracy and quality in the systems they develop and provide adequate consumer protection for NSW citizens.

Prohibit unexplainable, general AI or AI that exhibits signs of dangerous capabilities in critical infrastructure and government services

The NSW Government has an important role to play in mitigating the risks of AI to the public through safe and responsible use of AI in decision making and service delivery. The Commonwealth's debt assessment and recovery program which wrongly recovered debt programs using automated decision making[21] is an example of the misuse of AI and human oversight which led to irreparable physical and mental damage including lost lives.

What we are most concerned about is the potential for malfunction of very highly advanced AI in managing and operating critical infrastructure[22] in NSW, particularly with technologies based on deep learning which make decisions that its creators are unable to explain (the

[20] "Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)", The European Commission (2023).
[21] "Report of the Royal Commission into the Robodebt Scheme", Royal Commission into the Robodebt Scheme (Updated on 11 July 2023): The Royal Commission into the Robodebt Scheme has concluded. Commissioner Catherine Holmes AC SC presented the Report of the Royal Commission into the Robodebt Scheme to the Governor-General, His Excellency, General the Honourable David Hurley AC DSC (Retd) on 7 July 2023. It was tabled on 7 July 2023.
[22] Security of Critical Infrastructure Act 2018 (Aust Cth), section 9.

'black box' problem[23]). A single "jailbreak", "hallucination", or other malfunction of an AI system based on deep learning may cause us to lose control of essential services like dams, electricity and air traffic which could lead to large-scale, devastating consequences. AI developers do not understand how the models they develop work inside and they have yet to dependably demonstrate the safety and controllability of these models.

Critical infrastructure is a complex shared responsibility of the Commonwealth, states and local governments. We note the commencement of the RMP obligation on critical infrastructure in February 2023. In our submission to the Commonwealth's consultation on AI regulation earlier this year, we suggested updating the *Security of Critical Infrastructure Act 2018* to indefinitely prohibit the use of general-purpose AI (such as large language models), uninterpretable AI (e.g. based on deep learning), and "human-out-of-the-loop" AI applications. This will put Australia in line with other overseas jurisdictions that are already reviewing general-purpose and autonomous AI in critical infrastructure and defence[24].

We propose NSW prohibits use of the following technologies in service delivery, government administration or critical infrastructure:

- Overly powerful AI technologies or AI performing at the threshold of capability
- Technologies that exhibit signs of deception, self-awareness, situational awareness, inclinations to self-replicate, or other dangerous capabilities[25]
- Technologies that are specifically designed to facilitate any banned application
- Technologies that lack the degree of interpretability needed to verify that they are not being used for any banned application
- All applications that the draft EU AI Act lists in Title II[26] ("Prohibited AI practices", such as "subliminal techniques")
- Agentic ("human-out-of-the-loop") AI applications.

Identification of overly powerful AI technologies

It is not clear at what level of capability a model or system can be considered overly capable. We propose the NSW Government does not use AI technologies or applications that are more capable than the level of OpenAI GPT-3 or GPT-4 series of models. We initially pick GPT-3 as a benchmark because these models have been in existence since 2020. The specific threshold can be revised as new safety research is published.

[23] "Sparks of Artificial General Intelligence: Early experiments with GPT-4", Microsoft Research (22 March 2023): "elucidating the nature and mechanisms of AI systems such as GPT-4 is a formidable challenge that has suddenly become important and urgent" (page 95).
[24] Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023 (US).
[25] "Model evaluation for extreme risks", Toby Shevlane, et al. (24 May 2023).
[26] "Artificial Intelligence Act", European Parliament (Updated on 10 October 2023).

In practice that means banning the use of AI models that were trained using more than 10^{23} FLOP in compute (approximately the amount of compute used for the original GPT-3 175B)[27]. Generally, the larger the compute required to train the model, the more powerful it is. However, there is a possibility that future improvements in the algorithms and data quality will allow for the creation of more powerful AI using smaller computational power. We propose to give the responsible Minister the power to ban certain technologies as and when they are identified as dangerous.

Applying this threshold will not unduly limit benefits from generative AI and will not cause business disruptions because the vast majority of technologies are below this threshold, including most-used APIs from OpenAI (the current technology avant-garde).

We also propose that the NSW Government prohibit government use of AI that have been found to exhibit signs of dangerous capabilities of which we gave examples in "2. AI risks to NSW community":

- Cyber-offense
- Deception
- Persuasion and manipulation
- Political strategy
- Weapons acquisition
- Long-horizon planning
- AI development
- Situational awareness
- Self-proliferation.

There are legitimate and beneficial uses of other types of AI which can reduce costs, increase efficiency and improve the lives of NSW citizens in government service delivery, such as the use of computer vision to detect surface cracks in bridges and tunnels[28].

We consider the current state of AI capabilities to be at an optimal level where the benefits are maximised and risks are manageable. There are still great benefits to harness from the current level of capability for many years to come.

Develop guidance to NSW public sector on use of generative AI

Generative AI (AI that is used to create new text, images, video, audio or code) such as CHatGPT and Google's Bard are free to use and can be helpful in assisting public servants

[27] "Proposals", Stop AGI (2023).
[28] "Computer vision framework for crack detection of civil infrastructure—A review - ScienceDirect", Dihao Ai, Guiyuan Jiang, Siew-Kei Lam, Peilan He, Chengwu Li (21 October 2022).

to carry out their work provided sensitive information is not entered and output is checked for accuracy and bias. We propose guidelines in Appendix A to ensure safe and responsible use of these publicly available tools.

Redirect government funding in AI industry or state capability development to safety research

Australian businesses, including small businesses, already enthusiastically embrace AI due to its potential to automate their processes and improve profit margins[29]. Further industry support from the government is not required to grow the adoption of AI. But it may help mitigate the risks of the new technology. Therefore, we recommend redirecting existing industry funding towards AI safety research.

The NSW Government could provide assistance to develop a competitive advantage in technical AI safety research, a nascent field that is attempting to solve the alignment (ensure ‘values’ of increasingly powerful AI are aligned with human values[30]) and control problems of AI. Building knowledge in this field will have positive spillover effects on the development of safe and responsible AI (e.g. lower costs for new market entrants). There is currently little private incentive for the AI developers to prioritise responsible development and safety testing in the race to develop more advanced AI systems.

Government support could go towards:

- National standards institutes to work on means of quantitative assessment of AI capabilities and safety[31]; and
- Free up any existing research grants earmarked for AI capability research, computer science or fundamental science so they can be used in AI safety research (or establish new grants for AI safety research).

Additionally, if the NSW Government were considering support to help NSW businesses adopt AI, we suggest the funding be directed towards:

- Cybersecurity suppliers to review implementations of AI systems, such as compliance with emerging standards like OWASP Top 10 for Large Language Model Applications[32].
- AI ethics training businesses that can conduct community workshops for small businesses to teach best practices in compliance with the new AI regulations and principles, data protection, and related topics.
- AI ethics consulting businesses to review compliance with any new regulations and help businesses adjust to these requirements.

[29] “Australian retailers embrace the power of AI and automation”, Kaleah Salmon, eCommerceNews Australia (13 July 2023).
[30] “What is the AI alignment problem and how can it be solved?”, Edd Gent, New Scientist (10 May 2023).
[31] “Strengthening U.S. AI Innovation Through an Ambitious Investment in NIST”, Anthropic (April 2023).
[32] “OWASP Top 10 for Large Language Model Applications”, The OWASP Foundation (2023).

The funding should only go to local NSW or Australian companies (locally owned, with local employees). It should not be spent on credits for the use of AI tools and APIs. It should preferentially be given to support small businesses in higher-risk industries (e.g. local clinics that handle patients' health records and may begin to use AI in diagnosis or management of chronic conditions).

3. Comments on the NSW Government's AI Strategy, Ethics Policy and Assurance Framework

We support the actions in the AI Strategy to upskill public servants' AI knowledge and capabilities (see Appendix A for our proposed guidelines to use generative AI tools appropriately), and to preserve NSW citizens' right to question an AI-informed decision or outcome and understand the basis on which that decision was made and access to a review process. We support the mandatory principles in the Ethics Policy, they are broadly in line with those proposed or in use in international jurisdictions such as the UK, US, Canada, Singapore and the EU.

Add safety as a mandatory principle in the Ethics Policy

The Ethics Policy, AI Strategy and Assurance Framework (released in March 2022) predate the largely unanticipated arrival of powerful generative AI. They also predate the concerns leading AI industry experts have about "emergent abilities"[33] of large language models, including abstract thinking and dangerous capabilities (such as ability to suggest dangerous chemical compounds[34]) and whether they are truly capabilities or regurgitation of training data. The Center for AI Safety's overview of the main sources of catastrophic AI risks and growing examples which warrant serious concern[35].

Safety must be a fundamental principle guiding the use of AI by governments. AI, particularly large language models, is risky technology. The developers of the large language models that power Chat-GPT, Bard, DALL-E etc do not understand the inner workings of these models or how they make decisions. They are also inaccurate and we do not yet know how to control highly advanced AI systems. At the same time, AI capabilities continue to grow at an unprecedented rate.

What does safety look like?

The NSW Government should procure from AI developers that:

- continually test their models for safety (pre and post deployment of every new version or upgrade)
- disclose the results of independent safety evaluations (e.g. malicious use of the model and unintended consequences of use)
- disclose training data source
- devote a significant part of resources to AI safety research

[33] "Emergent Abilities of Large Language Models", Jason Wei, et al. (Updated on 26 October 2022): Ability is defined as "emergent if it is not present in smaller models but is present in larger models".

[34] "ChemCrow: Augmenting large-language models with chemistry tools", Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller (21 June 2023).

[35] "An Overview of Catastrophic AI Risks", Dan Hendrycks, Mantas Mazeika and Thomas Woodside (9 October 2023).

Safeguards should be in place to minimise potential risks such as misuse and malicious actors hacking into critical systems. The outcomes of automated decision systems need to be frequently monitored and verified to ensure they continue to meet their intended purpose, are robust, accurate and safe. Public servants need to be wary of claims of safety or 'safety washing', do due diligence and require companies to demonstrate with substantial evidence their technology is safe. We make suggestions on how to achieve this in the Assurance Framework.

Prohibit public entities from using high risk AI applications and technologies in provision of services or enforcement of powers

The NSW Government should prohibit the public sector and enforcement agencies from using facial recognition and biometric identification or any invasive mass surveillance technologies. They infringe on fundamental rights and can lead to disastrous consequences if misused or hacked into by malicious actors. Banning these applications is in line with the draft *EU AI Act*.

In the previous chapter, we recommended prohibiting use of highly advanced and dangerous AI technologies. This can be implemented through the Assurance Framework, for example by prohibiting entering into contracts with AI developers or sellers of such high risk technologies.

We suggest including the following clauses into contracts for the provision of AI to the NSW Government to give the government or purchaser the right to:

- notifications when the seller becomes aware of safety issues with the technology (e.g. dangerous behaviours or vulnerabilities) including newer versions of the model, when there is a lawsuit against the seller
- documentation regarding the sources of data used in model training and deployment, how the data is collected and maintained (this ensures transparency and accountability and can alert to the use of hateful content or personal information)
- retain ownership of government data inputs (i.e. not to be kept or owned by the seller and must not store the information or use it in future training runs)
- explanations of how the AI model makes decisions (in order to fulfil the right of NSW citizens to understand the basis on which an AI-informed decision is made)

We also propose to require NSW government agencies to include a statement of the use, inputs and a description of the operation of AI systems in annual reporting and in open access information to ensure transparency in government decision making and to allow for compliance monitoring by other agencies such as the Information and Privacy Commission NSW. This can be mandated through the GIPA Act.

4. Measures other jurisdictions, both international and domestic, that are relevant and adaptable to NSW

Impose transparency and disclosure requirements on supply of high risk AI systems

The Safety in Artificial Intelligence Act (SB 294) bill introduced in the California State Senate in September 2023 provides an example for NSW to influence the safe development of AI at the state level.

Being an intent bill, it is unable to go through the legislative process but nevertheless contains elements that we support. It targets the development of cutting edge, high-risk frontier AI systems (including large language models and generative AI that are highly capable) by[36]:

- requiring AI developers to disclose:
    - safety risks models are being tested for
    - actions and safety measures in response to warning signs of danger
    - information about the conditions in which adding unpredictable new capabilities would be dangerous
    - safety plans as they continue to scale to more advanced AI systems.
- Require commercial cloud computing providers to institute Know Your Customer policies for large-scale frontier model training runs
- establishing a review body would be established to address disclosures and conduct audits
- Establishing liability for insufficient measures taken to prevent misuse and unintended consequences that threaten public safety
- Create an AI cloud compute cluster dedicated to safety research.

Unlike California, NSW is not home to leading AI developers who carry out large-scale model training runs. However, we believe NSW can make a significant influence on the development of AI by imposing safety and transparency measures on AI developers (the likes of Google, Meta, Microsoft which all have offices and a significant business presence in NSW) when supplying their services to NSW customers. This approach is not new: for example, NSW imposes disclosures requirements for how businesses communicate with

[36] "Senator Wiener Introduces Safety Framework in Artificial Intelligence Legislation", Senator Scott Wiener, representing Senate District 11, San Francisco (2023).

customers before completing a sale[37] (section 47a of the Fair Trading Act NSW). The NSW Government could require AI developers to disclose information about the AI system's:

- performance characteristics
- training data used (with comprehensive references)
- intended context of use (to help businesses evaluate suitability of these systems to their context
- results of pre-deployment safety and performance evaluations.

We propose the above transparency requirements to apply to highly advanced AI systems with general capabilities as they are high risk. So this would not apply to low risk AI such as customer support chatbots, educational games and spam filters.

Restrict the supply of high risk AI systems into NSW to licensed AI developers

Some countries such as the UK, USA and EU are considering licensing schemes that restrict or place conditions on the development of highly capable and overly powerful AI and/or has safety conditions similar to those set out above.

An example is the Bipartisan Framework for the U.S. AI Act developed by Senator Richard Blumenthal & Senator Josh Hawley to establish a licensing regime overseen by an independent oversight body, liability for AI developers for privacy and civil rights violations and other harms, requiring disclosure of essential information about how AI systems work and other consumer protection measures[38].

Allowing only the supply of high risk AI systems from recognised overseas licensed developers is a practical and fast approach to protect NSW citizens without having to implement similar safety and transparency requirements.

Prohibit the development of increasingly powerful and intelligent general-purpose AI in NSW

Once overly powerful AI systems come into existence it is hard to prevent misuse and they can be copied. It is therefore prudent to prevent their creation in the first instance.

Currently, the vast majority of AI development (particularly sophisticated general-purpose models) takes place outside of Australia due to a variety of reasons such as access to computing requirements and skilled labour. The impact of this recommendation will therefore be minimal in NSW.

[37] "New disclosure obligations for NSW businesses", NSW Fair Trading, NSW Government (18 June 2020).
[38] "Blumenthal & Hawley Announce Bipartisan Framework on Artificial Intelligence Legislation", U.S. Senators Richard Blumenthal, D-CT and Josh Hawley, R-MO (09 August 2023).

Nevertheless, we think it is important to implement this recommendation because it responds to public warnings by AI computer scientists, researchers and leading AI developers of the profound risks to society and humanity as AI becomes more powerful[39].

We propose to identify and prohibit development that is more powerful than GPT-4[40] and to enforce this by monitoring compute in NSW via surveillance of power usage within data centres, mandate a know-your-customer (KYC) scheme and reporting of compute activities for cloud compute providers operating in NSW and impose tough penalties[41] for non-compliance.

[39] Statement on AI Risk | CAIS

[40] Pause Giant AI Experiments: An Open Letter - Future of Life Institute

[41] E.g. percentage-of-worldwide-turnover fines and criminal criminal penalties against corporations, their employees and directors, similar to sanctions under *EU GDPR* (GDPR Enforcement Tracker)

5. Impact of AI on jobs and the economy

People and businesses are concerned about job losses

AI models and tools are being used across the economy in a wide range of applications and playing an increasingly important role in productivity, growth and living standards.

Businesses in NSW use AI to replace people for some tasks, personalise customer service and create new products and services. NSW citizens interact with AI on a daily basis in search, automated assistants, social media and language services and we are becoming more aware of this interaction.

In a survey conducted by Roy Morgan[42] for the Campaign for AI Safety, 57% of Australians believe AI creates more problems than it solves and the most common reasons for agreeing with this statement were job losses and misuse of AI. PwC's 23rd Annual Global CEO Survey found that 29% of CEOs surveyed expressed worries that AI will displace more jobs than it creates[43]. The University of Queensland's 2023 Global Study found that most people (71%) disagree or are unsure that AI will create more jobs than it will eliminate. The World Economic Forum's May 2023 Future of Jobs survey globally found respondents expect structural job growth of 69 million jobs and a decline of 83 million jobs. This corresponds to a net decrease of 14 million jobs, or 2% of current employment[44]. Analysis commissioned by the UK Government is inconclusive about the net impact of AI on employment[45].

AI uptake across the economy is seen as a solution to NSW's slowing productivity growth

The NSW Productivity Commission is of the view that automation will not lead to widespread joblessness and that decreased demand for some occupations will be more than offset by increasing demand from other parts of the economy and new jobs created from the use of emerging technologies. It believes AI will revive NSW's slowing productivity growth and drive prosperity[46]. Similarly, the Commonwealth Productivity Commission states productivity gains can be significant, "from robot-assisted warehouses that automate online order fulfilment and reduce accidents, to AI-enabled IoT sensors installed in smart cities that allow real-time optimisation of infrastructure, energy and service use and maintenance notification"[47]. It does not comment on the impact of AI on employment.

[42] "Majority of Australians believe artificial intelligence (AI) creates more problems than it solves", Roy Morgan Research (29 August 2023).

[43] "PwC's 23rd Annual Global CEO Survey: Navigating the rising tide of uncertainty", PwC (15 Jan 2020).

[44] "Future of Jobs Report 2023", World Economic Forum (May 2023).

[45] "The potential impact of AI on UK employment and the demand for skills", Department for Science, Innovation and Technology and Department for Business, Energy & Industrial Strategy, GOV.UK (8 October 2021).

[46] "Adaptive NSW: how embracing tech could recharge our prosperity", NSW Productivity Commission, NSW Innovation and Productivity Council (November 2022).

[47] "Volume 4 - 5-year Productivity Inquiry: Australia's data and digital dividend", Australian Government Productivity Commission (7 February 2023): Report no. 100, page iv.

We question whether AI can drive prosperity in NSW and we are concerned about future standards of living as real wages remain stagnant across the country. An analysis of US wage bills has found a sharp slowdown in wage bill growth in the most recent 30 years than the previous four decades due to an acceleration of automation and a deceleration in the creation of new tasks[48].

In Australia, real wage growth has been falling for more than a decade and this trend precedes the COVID-19 pandemic. Average annual growth in real wages in the five years to November 2018 (0.5% per year) was significantly less than the average recorded in the previous five years to November 2013 (1.8% per year)[49]. In NSW, this trend exacerbated during the pandemic when real wages fell 2.5% between June 2021 and June 2022. One analysis found the past three years of falling real wages during the pandemic has led to the average wage now being able to purchase the same basket of goods as in 2009[50]. The factors behind this trend are not well understood.

We are concerned AI uptake could have significant redistributive wealth effects which may warrant government intervention to maintain standards of living

NSW workers who perform tasks that use AI and are hard to replace (inelastic labour supply) will likely benefit more from the adoption of these technologies than those in the opposite scenario (those working in tasks that can be performed by AI). It is likely that the number of NSW workers in the favourable scenario are few as they are likely to be very highly skilled, globally sought after workers. There is a chance increased adoption of AI and automation could lead to greater economic inequality (job loss and lower wages for workers and monopoly rents and profits for owners of AI technology which are Microsoft, Google, OpenAI[51], and Meta).

If the jobs displacement impact is large and laid-off workers are unable to find work in other parts of the economy, the NSW Government may need to recalibrate its taxation and welfare policies in order to maintain NSW citizens’’ standards of living. The costs of this may or may not be greater than the benefits of widespread adoption of AI.

[48] “Automation and New Tasks: How Technology Displaces and Reinstates Labor - American Economic Association”, Daron Acemoglu and Pascual Restrepo (2019): Journal Of Economic Perspectives, Vol. 33, No. 2, pp. 3-30.

[49] “The extent and causes of the wage growth slowdown in Australia”, Geoff Gilfillan, Statistics and Mapping Section, Parliament of Australia (9 April 2019).

[50] “It’s good news that real wages are no longer falling – but the fall has already been deep”, Greg Jericho, The Guardian (17 August 2023).

[51] 50% owned by Microsoft.

AI technology is supplied by a handful of very powerful businesses which could negatively impact consumer welfare and competition across the economy

We are concerned that AI’s increased use in all parts of the economy could make AI developers (dominated by Google, Microsoft, Meta) more powerful than they currently are and have broad influence over the economy and our lives and livelihoods.

AI can exhibit significant economies of scope and, as it is increasingly being used across multiple markets, this could lead to these businesses becoming de facto conglomerates. These businesses could command high prices, lower the quality of goods or services, cause consumer detriment and make entry by new businesses very difficult, reducing competition in NSW and elsewhere. There would also be significant implications for all of us who interact with AI on a daily basis (e.g. user manipulation, false and misleading information, preventing consumers from making informed choices about products and services) if these companies were to misuse their market power. The Competition and Markets Authority, the UK’s regulator for competition and consumer protection, shares this concern[52]. AI developers such as Anthropic reportedly want to “automate large portions of the economy”[53]. There is empirical evidence that AI investment is associated with increased market concentration, and higher AI adoption and larger gains from AI investments for larger companies[54].

The ACCC’s investigation into digital markets has found that Big Tech companies (the same ones that supply AI technology) leverage market power into downstream markets using their online platform and giving preferential treatment to third party suppliers and to themselves[55].

[52] “CMA response to DCMS pro-innovation approach for regulating AI”, presented to Parliament by the Secretary of State for Digital, Culture, Media and Sport by Command of Her Majesty (Updated 20 July 2022).

[53] “Anthropic’s $5B, 4-year plan to take on OpenAI”, Kyle Wiggers, Devin Coldewey, Manish Singh, TechCrunch (7 April 2023).

[54] “Artificial Intelligence, Firm Growth, and Product Innovation”, Tania Babina, Anastassia Fedyk, Alex Xi He and James Hodson (Updated on 3 February 2023): Journal of Financial Economics (JFE), Forthcoming.

[55] “Digital Platforms inquiry 2017-19”; “Digital platform services inquiry 2020-25”; and “Digital Advertising Services inquiry 2020-21”, Australian Competition and Consumer Commission.

Appendix A: Proposed generative AI guidelines for the NSW public service

Currently only the NSW Department of Education has guidelines on the use of generative AI. We recommend extending these guidelines to the rest of the public sector and making them mandatory to comply.

We propose the following which can be implemented without delay:

1. Public servants and contractors must not enter private or sensitive information into generative AI tools such as Chat-GPT, Bard, DALL-E, etc. because the information is often transferred overseas and may be used for model training purposes (i.e. be permanently incorporated into AI models under the control of foreign actors).

2. In the writing of policy documents, drafting legislation or other forms of legal writing, public servants and contractors should use AI software with transparent training datasets. This is due to the possibility that the biases[56][57] in the training data can sway the thinking of the writers as they use autocomplete functionality.

3. Public servants should be made familiar with the pitfalls of existing AI technologies, such as "hallucinations"[58].

4. AI tools that are based on deep learning (including most generative AI) are non-transparent black boxes and therefore must not be used for any form of ADM.

5. Public servants should adhere to *Australia's AI Ethics Principles*.

We suggest including these guidelines in onboarding materials for new staff and annual refresher training.

Breaches of this guidance should have the same sanctions as those for breaches to the NSW Public Sector Code of Conduct (suspension, reduction in classification, reassignment of duties, termination of employment, etc).

[56] "The politics of AI: ChatGPT and political bias", Jeremy Baum, John Villasenor (8 May 2023).

[57] "Political Bias in Large Language Models", Lucas Gover (17 May 2023): The Commons: Puget Sound Journal of Politics: Vol. 4: Iss. 1, Article 2.

[58] "Hallucination (artificial intelligence)", Wikipedia (16 July 2023).