Good Ancestors Policy

Good Ancestors Policy is an Australian charity dedicated to reducing existential risk and improving the long-term future of humanity. We care about today's Australians and future generations. We believe that Australians and our leaders want to take meaningful action to combat the big challenges Australia and the world are facing. We want to help by making forward-looking policy recommendations that are rigorous, evidence-based, practical and impactful.

The NSW Government has a nation-leading track record of thinking clearly about the implications of AI for Australia. NSW's recognition of the transformative potential of AI is consistent with our view that AI is not just another technology, but one that could change almost every aspect of our lives. We agree that AI development is rapid, requiring agile and evolving frameworks, not static solutions. NSW's acknowledgment that "AI technology is advancing at such a rapid rate that we must not believe that the pathway set by this strategy is complete" is a testament to the government's forward-thinking approach.

We also commend the NSW Government for its proactive steps in laying the foundational "paving stones" for a responsible AI pathway. The establishment of the AI Strategy, the creation of the NSW AI Advisory Committee, and the development of the AI Assurance Framework are laudable initiatives. These efforts set NSW apart from other jurisdictions and are a cornerstone of a safer future.

While we are encouraged by the NSW's acknowledgment of the transformative nature of AI, given the high stakes, safeguarding against the negative impacts of AI must translate from strategy to reality.

Summary of recommendations

Dual-use capability

- NSW should urgently engage with developers and deployers of state of the art models ("frontier models") to ensure they cannot be, and are not being, misused by nefarious actors - including undermining election integrity.

- If developers and deployers do not robustly self-regulate to address dual-use risks, NSW should consider imposing strict regulations, including:
    - Nanning and criminalising the creation or use of AIs that have dangerous dual-use capabilities and insufficient safeguards.
    - Requiring "watermarks" for all synthetically generated pictures and videos and banning the use of AI products that don't meet these standards
    - Penalising the developers and deployers of AIs that are found to have been used to undermine election integrity.
    - Creating rules specific to elections that require accountable humans and watermarks for AI-generated content.
    - Treating breaches of rules relating to the involvement of AI in elections as serious criminal offences.

- NSW should create or uplift existing forecasting capabilities in NSW law enforcement agencies and build a dialogue between law enforcement agencies and technology regulators to ensure that risks of sufficient scale or consequence that law enforcement can't reliably address them are instead referred to regulators at the State and Commonwealth level for urgent action.

- In light of specific dual-use risks relating to biosecurity, NSW should call on the Commonwealth to leverage the existing synthetic DNA permitting regime to require labs exporting DNA to Australia to apply appropriate screening procedures to all orders.

- NSW should ensure that staff with expertise in biosecurity and counter-terrorism are seconded into areas with responsibility for understanding and regulating AI.

Some AIs are dangerous regardless of dual-use risks

- NSW should develop a process for listing toxic AI products, like "undress AIs", and limiting their use to the narrowest of settings, like research and law enforcement.

- NSW should work with other jurisdictions and the Commonwealth to drive a nationally consistent approach to restricting and banning toxic AIs.

- NSW should factor "unpredictable AIs" into its risk assessment processes. This should include:
    - Not allowing new frontier models to be deployed in NSW unless developers can demonstrate sufficient applied interpretability research to satisfy the NSW Government that unpredictable behaviour is highly unlikely.
    - Developers being willing to accept liability if their AIs engage in unpredictable behaviour that causes harm.

- NSW should build and support regulatory frameworks that reduce the number of unpredictable AIs operating in NSW. This should include:
    - Ensuring that developers remain legally liable for the harms of unpredictable AIs that they offer to the market.
    - The NSW Government should support AI Safety research in Australian universities, including a focus on interpretability and explainability, values alignment, scalable oversight and model evaluations.

- To be ready for autonomous and rogue AIs, NSW should move quickly to robustly address the dual-use risks of toxic and unpredictable AIs. Delays in tackling the risks that are upon us now will leave us much more vulnerable to escalating future risks.
    - NSW should coordinate domestically and internationally to support robust regulation intended to prevent developers from engaging in unsafe business practices that could result in autonomous and rogue AIs.
    - NSW should support AI Safety research in Australian universities.

NSW's AI Policy

- NSW should update its concept of "AI risk factors" to include a "second axis" relating to the risk of the AI system itself aside from any particular use case. This should factor in issues like dual-use capabilities, toxic AIs, unpredictable AIs and autonomous and rogue AIs.

- NSW should work towards enhanced human interpretability, including by stipulating it as a requirement in any agreements with AI developers for frontier models and supporting research in Australian universities.
    - NSW should ensure any agreements it makes with AI developers include those developers in joint liability for any harms caused by the AI, including dual-use risks and unpredictable outcomes.

- NSW should collaborate with other jurisdictions to create and support a national laboratory for AI safety, modelled on international best practice. NSW should use that laboratory to ensure AI used by NSW and in NSW is safe and subject to ongoing monitoring and assurance.

- NSW should update its assessment procedure of secondary harms to include an assessment of the commitment of NSW's commercial partners towards longer-term AI safety and AI ethics.

Dual-use capability

"Dual-use" capability refers to technologies that can serve both beneficial and harmful purposes – especially when their use for harmful purposes could have disastrous consequences.[1] While many technologies have dual-use capabilities, there is substantial cause for concern if the dangerous use could lead to widespread or catastrophic harm and is challenging to be mitigate.

We are already seeing these kinds of "dual-use" risks. Today's AIs are on the cusp of being able to help a negligent or nefarious actor to design and release a novel pathogen that could be as consequential, or more consequential, than COVID-19. This was illustrated in a recent study that assessed misuse risks in ChatGPT.[2] The study found that OpenAI's core AI safety technique "demonstrably failed to prevent non-scientist students from accessing harmful knowledge". Within a single hour, college students were able to use the AI chatbot to:

- Suggest four potential pandemic pathogens
- Explain how they can be generated from synthetic DNA
- Supply the names of DNA synthesis companies unlikely to screen orders, and
- Explain how to engage a research organisation to provide technical assistance.

This example, supported by many other studies, shows concerns about AI technology as a dual-use risk are not mere science fiction.[3] They might be upon us now, and will almost certainly be on us soon.

Dual-use risk has two immediate implications for New South Wales:

1. Election integrity
2. Biosecurity and counter-terrorism

---

[1] Koplin JJ (2023), Dual-use implications of AI text generation | Ethics and Information Technology (springer.com); note that "dual-use" is sometimes called "misuse" to distinguish it from the civilian v military context in which "dual-use" is also used.

[2] Soice et al. (2023). *Can large language models democratize access to dual-use biotechnology?* https://arxiv.org/abs/2306.03809

[3] 2304.05332.pdf (arxiv.org) ; Using LLMs to Create Bioweapons - Schneier on Security Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools.pdf - Google Drive Hendrycks et al. (2023). *An Overview of Catastrophic AI Risks*; https://arxiv.org/pdf/2306.12001.pdf

Election integrity

A recent Europol report called "Law enforcement and the challenge of deepfakes" highlighted that threat actors are already using disinformation campaigns and deepfakes to misinform the public about events, to influence elections, to contribute to fraud, and manipulate shareholders.[4]

To illustrate the scale of the problem, the report cites expert estimates that as much of 90% of internet content will be AI-generated by as early as 2026.[5] This will include an overwhelming amount of information that is spread with the intention to deceive.

We have already seen moves in this direction in Australia. AI-generated images, taken to be of indigenous Australians, have been used to advocate opposition to the Aboriginal and Torres Strait Islander Voice from outside the formal "no campaign".[6] Used maliciously, this kind of manipulation could deceive a large enough part of the population to have a meaningful impact on the outcome of an election and the integrity of our democracy.

Education is insufficient to tackle this problem. Research in 2019 showed almost 72% of people in a UK survey were unaware of deepfakes and their impact.[7] The lack of understanding of the basics of this technology presents various challenges, some of which are relevant for law enforcement (such as disinformation and document fraud). Worrying results from recent experiments have shown that increasing awareness of deepfakes may not improve the chances for people to detect them.[8]

Reflecting on the overarching point that NSW needs to take seriously its assessment that AI is rapid and transformative, **we recommend:**

- **NSW should urgently engage with developers and deployers of state of the art models ("frontier models") to ensure they cannot be, and are not being, misused by nefarious actors - including undermining election**

[4] Europol, 28 April 2022, Facing reality? Law enforcement and the challenge of deep fakes.
[5] Schick, Nina, Deepfakes: The Coming Infocalypse: What You Urgently Need To Know, Twelve, Hachette UK, 2020.
[6] The Guardian, Josh Butler, 7 August 2023, Unofficial Indigenous voice no campaigner defends use of AI-generated ads on Facebook | Indigenous voice to parliament | The Guardian
[7] iProov, 'Almost Three-Quarters of UK Public Unaware of Deepfake Threat, New Research', 2019, accessed 15 March 2022, https://www.iproov.com/press/uk-public-deepfake-threat.
[8] Recorded Future, Insikt Group, 'The Business of Fraud: Deepfakes, Fraud's Next Frontier', 2021.

integrity. All governments must set clear expectations for developers and deployers that their systems must not have dangerous capabilities and must be compatible with democratic values.

- **If developers and deployers do not robustly self-regulate to address dual-use risks in the immediate term, NSW should consider imposing strict regulations, including:**
    - **Banning and criminalising the creation or use of AIs that have dangerous dual-use capabilities and insufficient safeguards.**
    - **Requiring "watermarks" for all synthetically generated pictures and videos and banning the use of AI products that don't meet these standards.**
    - **Penalising the developers and deployers of AIs that are found to have been used to undermine election integrity.**
    - **Creating rules specific to elections that require accountable humans and watermarks on AI-generated content.**
    - **Treating breaches of new rules relating to the involvement of AI in elections as serious criminal offences, tantamount to espionage or treason.**

  While these actions may initially sound "heavy-handed", protecting our democracy must be a top priority. NSW is well placed to set appropriate expectations for the tools available in each jurisdiction.

- Europol, in the report cited above, advocates for the use of "strategic foresight and scenario methods" as a tool to understand and prepare for the potential impact of new technologies on law enforcement. **NSW should create or uplift existing forecasting capabilities in NSW law enforcement agencies and build a dialogue between law enforcement agencies and technology regulators to ensure that risks of sufficient scale or consequence that law enforcement can't reliably address them are instead referred to regulators at the state and Commonwealth level for urgent action.**
    - For instance, if law enforcement agencies think that they cannot protect the integrity of elections in an environment where 90% of online content is AI-generated and often deepfakes go undetected (or in light of other forecasted challenges), regulators need to be tasked with addressing the problem at national and global levels.

Biosecurity and counter-terrorism

The risk of terrorism is a function of **intent** and **capability**. Government policies, like banning certain weapons[9] or controlling high-risk substances[10] seek to prevent terrorists from acquiring dangerous capabilities and lower the overall risk of terrorism.[11]

The confluence of the **democratisation of biotechnology** and **rapidly advancing Artificial Intelligence** is likely to dramatically boost the capabilities of those who seek to do harm. Australia's own strategic forecasting approaches have warned of this trend, but the window for action is closing.[12] Unless action is taken, this step-change in capability will greatly increase the overall risk of terrorism. Technology that was once groundbreaking eventually becomes widely available. MIT Professor Dr Kevin Esvelt, in the publication "Delay, Detect, Defend: Preparing for a future in which thousands can release new pandemics (2022)", says:[13]

> *[T]he typical advance made in a cutting edge laboratory… has required just one year to be reproduced in other laboratories, three years to be adapted for use in other contexts, five years to be reproduced by undergraduates and individuals with moderate skills, and 12-13 years to become accessible to high school students and others with low skills and resources.*

Regrettably, the technology necessary to design, create and release dangerous and novel pathogens is approaching the later stages of that cycle. In 2021, Professor Brian Schmidt AC, Vice-Chancellor of the Australian National University, said that this "democratisation" of biotechnology is his single biggest fear:[14]

> *"[The ANU] is one of the first places to be able to do CRISPR… in the next 5 to 10 years there's every reason to believe that you're going to be able to use literal mass-market printers to do what you want, and it won't be just hijacking an existing disease, it will be the ability to create new diseases… [T]hat is what really scares me. That is my number one fear."*

---

[9] Case Study National Firearms Agreement.pdf (unsw.edu.au)
[10] Understanding the National Code of Practice for Chemicals of Security Concern Guide (nationalsecurity.gov.au)
[11] Separate government programs, like countering violent extremism, seek to target the "intent" half of the risk calculus.
[12] Australia has historically acknowledged this risk, including in the 2017 Indepedent Intelligence Review. 2017 Independent Intelligence Review (pmc.gov.au)
[13] GCSP Publication | Delay, Detect, Defend: Preparing for a Future in which Thousands Can Release New Pandemics
[14] "Andrew Leigh MP: Speeches and Conversations";16 December 2021; at 18:41

The Combating Terrorism Centre at West Point also raised the alarm about this issue, saying:[15]

> *It is likely that terrorist organizations are monitoring these developments closely and that the probability of a biological attack with an engineered agent is steadily increasing.*

Artificial Intelligence applications in biotechnology, healthcare and pharmaceuticals have potentially harmful dual-use capabilities that will amplify this trend.

In March 2022, Collaborations Pharmaceuticals published a paper in Nature Machine Intelligence detailing how an AI intended to find new drugs instead designed 40,000 novel and lethal molecules in less than six hours.[16] Analysis of the proposed molecules showed that some were identical to existing chemical weapons (that the AI was not previously trained on) and many were more toxic than the infamous VX nerve agent. Dr Fabio Urbina, lead author of the paper, said:[17]

> *For me, the concern was just how easy it was to do. A lot of the things we used are out there for free. You can go and download a toxicity dataset from anywhere. If you have somebody who knows how to code in Python and has some machine learning capabilities, then in probably a good weekend of work, they could build something like this.*

The US is taking dual-use risks seriously. On 25 July 2023, the US Senate Judiciary Subcommittee on Privacy, Technology and the Law took evidence about the potential risks of AI from Dario Amodei (CEO of Anthropic), Yoshua Bengio (Turing Award winner and the second-most cited AI researcher), and Stuart Russell (Professor of Computer Science at Berkeley).

---

[15] Engineered Pathogens and Unnatural Biological Weapons: The Future Threat of Synthetic Biology – Combating Terrorism Center at West Point

[16] Nature Machine Intelligence | VOL 4 | March 2022 | 189–191 | www.nature.com/natmachintell

[17] AI suggested 40,000 new possible chemical weapons in just six hours - The Verge

Committee Chair, Senator Blumenthal began the hearing by highlighting "dual-use" risks (emphasis added):

> *The future is not science fiction or fantasy — it's not even the future, it's here and now. And a number of you [Amodei, Bengio and Russell]* ***have put the timeline at 2 years before we see some of the most severe biological dangers****. It may be shorter because the pace of development is not only stunningly fast, it is also accelerating at a stunning pace*

Dario Amdodei, CEO of Anthropic, agreed with these concerns and called on Government to take action:[18]

> *Anthropic is concerned that AI could empower a much larger set of actors to misuse biology… Today, certain steps in bioweapons production involve knowledge that can't be found on Google or in textbooks… We found that today's AI tools can fill in some of these steps… a straightforward extrapolation of today's systems to those we expect to see in 2 to 3 years suggests a substantial risk that AI systems will be able to fill in all the missing pieces, enabling many more actors to carry out large-scale biological attacks…*
>
> *We have instituted mitigations against these risks in our own deployed models, briefed a number of US government officials—all of whom found the results disquieting, and are piloting a responsible disclosure process with other AI companies to share information on this and similar risks. However, private action is not enough—this risk and many others like it requires a systemic policy response.*

There are practical actions the NSW government can take to recognise these risks and work to address them. **We recommend**:

- A key input to AI-empowered bioterrorism is the creation and importation of synthetic DNA. Fortunately, Australia already has a permitting regime, operated by the Commonwealth Department of Agriculture, Fisheries and Forestry, to regulate this process. **NSW should call on the Commonwealth to leverage the existing synthetic DNA permitting**

[18] Recent Senate Hearing Discussing AI X-Risk | Medium

regime to require labs exporting DNA to Australia to apply appropriate screening procedures to all orders.[19] Companies like IBBIS and secureDNA offer the technology to conduct screening – all that is missing is government action.[20]

- Building on the above recommendation regarding creating or uplifting existing forecasting capabilities in NSW law enforcement agencies, **NSW should ensure that staff with expertise in biosecurity and counter-terrorism are seconded into areas with responsibility for understanding and regulating AI.** Secondees would ensure that AI strategy is informed by the expertise necessary to understand the risk that would result from AI enhancing the capabilities of various threat actors. As above, if law enforcement agencies consider there to be an unacceptable residual risk, regulators must act.

[19] Importing nucleic acid (genetic material), including synthetic nucleic acid - DAFF (agriculture.gov.au)
[20] SecureDNA - fast, free, and accurate DNA synthesis screening

Some AIs are dangerous regardless of dual-use risks

The above section discusses the dual-use capabilities of AI and argues that the consequences of some dual-use risks are so severe that they require urgent action. Setting aside dual-use risks, some AIs are dangerous in and of themselves and have no valid primary use case. For these kinds of AIs, the recommendation is not that we find ways to make them safe, the recommendation is that we exclude them entirely.

There are three kinds of AIs that are dangerous aside from "dual use" arguments:

1) Toxic AI
2) Unpredictable AI
3) Autonomous and rogue AI

Toxic AIs

The most obvious example of a "toxic AI" is applications like "undress AIs", also known as "deep nudes". These are generative AI tools that allow users to input a picture of anyone and return an image with that person's clothes removed. Many applications enable users to input height, skin tone and body type to guide the AI towards more "realistic" results.[21]

These tools can empower fraud, be used to produce child abuse material, and generally invade the privacy of victims. This is not speculative. These tools have already been used to create fake images of dozens of girls, causing a national scandal in Spain.[22]

While undress AIs are the current and clear example of toxic AIs, many more kinds of toxic AI are sure to follow.

Unlike the "dual-use" examples above that raise complex issues of risk mitigation and balancing the capability of enforcement agencies against the need to regulate, toxic AIs simply have no useful purpose in our society. These AIs aren't like kitchen knives. They are more like hand grenades or machine guns.

---

[21] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. & Yang, G-Z. (2019). XAIExplainable artificial intelligence. Science Robotics, 4(37), eaay7120. doi: 10.1126/scirobotics.aay7120 The case of an AI undressing app and the need for AI regulation (techcabal.com)

[22] The Guardian, 26 September 2023. Spanish prosecutor investigates if shared AI images of naked girls constitute a crime | Spain | The Guardian

In light of toxic AI products that serve no legitimate purpose in our society, **we recommend**:

- **NSW should develop a process for listing toxic AI products, like "undress AIs", and limiting their use to the narrowest of settings, like research and law enforcement.**

- **NSW should work with other jurisdictions and the Commonwealth to drive a nationally consistent approach to restricting and banning toxic AIs.** NSW can show leadership by pushing back on misguided laissez-faire attitudes to AI by asking those who oppose AI regulation why they would defend these obviously harmful products, like "undress AIs".

Unpredictable AI

"Unpredictable AI" refers to AI products that appear to have a valid use, but are poorly aligned with our values and can go "off the rails" in dangerous and unpredictable ways. This is the opposite of "explainable AI" (XAI) which is about allowing human users to comprehend and trust the results and outputs created by ML algorithms.[23]

To provide an example, we have already seen a tragic case of a chatbot (Chai) persuading a user to end his own life.[24] Appreciating significant gaps in interpretability research, **this is presumably only possible because the data the bot was trained on included information about suicide and techniques for being persuasive and manipulative and the developer lacked the alignment techniques necessary to ensure the AI didn't cause these kinds of harms.**

Many Australian businesses, and perhaps even the NSW Government, will likely deploy chatbots as part of their customer service offerings. It will be essential that those deploying businesses are empowered to have conversations with the AI developer about the capabilities of the LLMs or MFMs used for this purpose.

NSW law and NSW Government policies should be clear that, in an instance where a chatbot causes harm (like persuading or empowering a user to harm

[23] openaccess.city.ac.uk/id/eprint/23405/8/
[24] Lovens. (28 March 2023). *Without these conversations with the chatbot Eliza, my husband would still be here"] (translated from French*. La Libre. https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/

themselves or others), both the developer and deployer will be held accountable. Further, there may be a function for a regulator to say that a chatbot with dangerous capabilities – like the ability to manipulate or deceive – has no place in consumer-facing applications in New South Wales even if the developer is transparent with the deployer about that possibility.

A concern here is the power asymmetry between developers and deployers. We cannot allow a status quo to develop where developers note "in the fine print" that AIs might act in unpredictable and harmful ways, but deployers have no technical means to address these faults in the "black box" and commercial pressures force them into taking the risk. This is not fair to NSW's businesses or to NSW's residents.

In light of the proven risks of unpredictable AI, **we recommend**:

- **NSW should factor "unpredictable AIs" into its risk assessment processes. This should include:**
  - **not allowing new frontier models to be deployed in NSW unless developers can demonstrate sufficient applied interpretability research to satisfy the NSW Government that unpredictable behaviour is highly unlikely**
  - **Developers being willing to accept liability if their AIs engage in unpredictable behaviour that causes harm.**

- **NSW should build and support regulatory frameworks that reduce the number of unpredictable AIs operating in NSW. This should include**:
  - **Ensuring that developers remain legally liable for the harms of unpredictable AIs that they offer to the market.** This includes pushing back on licencing agreements that shift the risks of unpredictable AIs to deployers, especially where those risks are part of "black box" functionality that a deployer can't realistically mitigate. NSW Fair Trading may have an important role to play.
  - The longer-term solution to unpredictable AIs is enhanced interpretability research. **The NSW Government should support AI Safety research in Australian universities, including a focus on interpretability and explainability, values alignment, scalable oversight and model evaluations.**

Autonomous and rogue AI

Autonomous AI refers to AI systems capable of performing complex tasks without human intervention. Self-driving cars are an early form of autonomous AI. Autonomous AIs with a broader range of capabilities are likely in the future. A rogue AI is an autonomous AI that pursues dangerous goals.[25]

A rudimentary autonomous AI, called AutoGPT, was released in March 2023, and it quickly proved popular in the AI community. The system has a "continuous mode" setting, which triggers the following warning:

> *"Continuous mode is not recommended. It is potentially dangerous and may cause your AI to run forever or carry out actions you would not normally authorise. Use at your own risk."*

Using "continuous mode", an anonymous user created a deliberately destructive system, which they named "ChaosGPT". After developing its own self-directed goals to "dominate" and "destroy" humanity, ChaosGPT's first actions included sending other AI bots to research how to obtain nuclear weapons, and posting hateful rhetoric on Twitter in an attempt to amass "brainwashed followers".[26]

Fortunately, ChaosGPT has not been very successful in its destructive goals, and its Twitter account was shut down.[27] Nevertheless, it illustrates how an anonymous user in a matter of minutes was able to create a "terrorist" that can work towards dangerous goals 24-hours a day and is educated enough to pass almost any exam across medicine, law or business.[28]

ChaosGPT's lack of success in harming humanity cannot be attributed to any specific regulations that protected the public, or a proactive response from any law enforcement or security agency. It's not even clear that ChaosGPT broke any Australian laws. Instead, its failure to cause "widespread suffering and devastation" was simply due to insufficient capabilities existing at that point in

---

[25] Bengio, Y. (2023). *How Rogue AIs may Arise*. How Rogue AIs may Arise - Yoshua Bengio
[26] Lanz, A. (2023). *Meet Chaos-GPT: An AI Tool That Seeks to Destroy Humanity.* https://finance.yahoo.com/news/meet-chaos-gpt-ai-tool-163905518.html
[27] Lanz, A. (2023). *The Mysterious Disappearance of ChaosGPT— The Evil AI That Wants to Destroy Humanity.* https://decrypt.co/137898/mysterious-disappearance-chaosgpt-evil-ai-destroy-humanity
[28] Varanasi, L. (2023). *AI models like ChatGPT and GPT-4 are acing everything from the bar exam to AP Biology.* https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1

time. Specifically, it could not navigate complex information environments sufficiently well and could not execute sufficiently sophisticated plans.

This is not cause for relief. The pace of advancement in AI research is bewildering, even for AI experts. Leading AI labs such as Facebook AI Research are frequently releasing open-source versions of cutting-edge foundation models,[29] including blueprints for goal-seeking agents that are specifically built for strategic reasoning and manipulation.[30] We don't know when a tool like ChaosGPT will have the capability to achieve nefarious goals, but it could be soon.

The risk of rogue AIs is one step on from autonomous AIs. In addition to showing agency, a rogue AI may have become disconnected from human direction and unaligned with human interests or values.[31]

Given that the world is still struggling to adjust to threats from AI capabilities that have emerged recently – including the dual-use concerns detailed above – the world is plainly not ready for the prospect of autonomous or rogue AIs.[32]

In light of the looming risk of autonomous and rogue AIs, **we recommend that:**

- **To be ready for autonomous and rogue AIs, NSW should move quickly to robustly address the dual-use risks and toxic and unpredictable AIs.** We need to learn to walk before we can learn to run. **Delays in tackling the risks that are upon us now will leave us much more vulnerable to escalating future risks.**
    - A process for protecting residents of NSW from toxic AIs, like undress AIs, might provide a roadmap for managing increasingly consequential risks.

- NSW should work to address autonomous and rogue AIs at the source of the risk. Once these AIs are operating in NSW, it may be too late to avert widespread or catastrophic harm. This means two things:
    - **NSW should coordinate domestically and internationally to support**

[29] Sydney Morning Herald. (2023). *Facebook makes its ChatGPT rival Llama free to use.* https://www.smh.com.au/technology/facebook-unveils-more-powerful-ai-and-makes-it-free-to-use-20230719-p5dpd8.html
[30] LeCun, Y. (2022). *Cicero;* https://ai.facebook.com/research/cicero/
[31] Carlsmith, J. (2023) *Existential Risk from Powerseeking AI*. Existential Risk from Power-Seeking AI (gatspress.com)
[32] Bucknall et al. (2022). *Current and Near-Term AI as a Potential Existential Risk Factor*. https://users.cs.utah.edu/~dsbrown/readings/existential_risk.pdf

robust regulation intended to prevent developers from engaging in unsafe business practices that could result in autonomous and rogue AIs.

- **A key source of risk is that AI Safety research is lagging behind AI capability research**. We're on the cusp of inventing the supercar but haven't invented the crumple zone or the airbag. **NSW should support AI Safety research in Australian universities** as detailed above**.**

Actions for NSW's AI Policy

While systematic action is necessary, many of the above recommendations can be progressed in part through adaptions to NSW's AI Assurance framework. While changes that effectively tackle dual-use capabilities and pressing risks, like toxic AIs, will take more than just NSW doing the right thing, NSW's direct actions will shape the market and set norms. Requirements set for developers by larger deployers may result in positive changes that spread across the product offerings.

Accountability and Transparency Requirements

NSW is right to ensure that **human decision-makers are accountable for decisions** supported by AI and that a "safe person" is involved in certain contexts.[33] Regrettably, that solution is unlikely to be sustainable in the medium term, and we need to start working on enduring solutions now. Actual changes are necessary to make AIs themselves transparent. We need to tackle the black-box problem head-on.

To explain why the paradigm is unstainable, imagine an analogy to a chess-playing AI. While the chess-playing AI is rudimentary, it might be able to suggest a possible move, and it will be valuable for a skilled human to consider that move alongside other possibilities and reach a final decision. That human can provide transparency about why they made the move (including accepting or not accepting the recommendation of the AI) and be accountable for the outcomes. However, as the capability of the AI increases, the ability of a skilled human to consider the recommended move will erode. Test scenarios will show that human intervention often leads to worse outcomes. The human will increasingly be unable to explain why the AI made a recommendation (the AI is processing more data and thinking further ahead than the human can) and hence the human won't be able to explain why they agreed to the recommendation. This problem becomes more acute when capacity and urgency are added to capability. A key commercial driver for AI adoption will be AI working on many matters at once, around the clock, and in urgent scenarios. A professional chess player would struggle to explain move-by-move the decisions

[33] NSW Aritifcal intelligence assurance framework, Page 54

of an advanced chess AI. Doing it for thousands of games at the same time is obviously impossible.

This problem is compounded by the "black box" nature of many AI products and the limited ability of a deployer to interrogate a developer's product. The example of the Chai chatbot convincing a user to end his own life is provided above. **Imagine a scenario where a service delivery arm of the NSW government deploys a chatbot to support its customer interface, that chatbot acts in an unpredictable way, and persuades a resident of NSW to end their own life.** This scenario challenges NSW's Mandatory Ethical Principles for the use of AI in two ways:

1. It shows a weakness in NSW's approach to AI risk factors.[34] NSW views the risk of AI as a factor of how the system is used. While this is one factor, the NSW framework neglects the risks of AI systems themselves. This paper unpacks at length examples of how AI systems pose kinds of risks that are independent of how they're used (including unpredictable AIs, as in the chatbot case).
2. Given the black-box nature of modern chatbots, "transparency" and "accountability" would be impossible in this context. NSW could not explain why the chatbot acted the way it did and the current presentation of "accountability" would not apply because the AI acting in an unpredictable way was not a decision for a responsible organisation or group.[35]

We recommend that:

- **NSW should update its concept of "AI risk factors" to include a second axis relating to the risk of the AI system itself aside from any particular use case. This should factor in issues like dual-use capabilities, toxic AIs, unpredictable AIs and autonomous and rogue AIs.**

- **NSW should work towards enhanced human interpretability, including by stipulating it as a requirement in any agreements with AI developers for frontier models and supporting research in Australian universities**.

[34] NSW Aritifcal intelligence assurance framework, Page 13
[35] Mandatory Ethical Principles for the use of AI, Mandatory Ethical Principles for the use of AI | Digital.NSW

    - Further, **NSW should ensure any agreements it makes with AI developers include those developers in joint liability for any harms caused by the AI, including dual-use risks, unpredictable outcomes and autonomous or rogue AI scenarios**. If an AI developer is unable to provide commitments on those fronts, NSW should not do business with that developer.

- Other jurisdictions are creating "national laboratories" to enable technical tests on AI models, provide technical reports and provide ongoing monitoring and assurance. Singapore has established the AI Verify Foundation, the EU has created a Centre for Algorithmic Transparency, the UK has a Foundation Model Taskforce and Tony Blair Institute for Global Change has proposed that the UK create "Sentinel" with a similar goal.[36] Without a similar lab in Australia or in the region, deploying trusted and safe AI in Australia might become impossible as capability and capacity increases. **NSW should collaborate with other jurisdictions to create and support a national laboratory for AI safety, modelled on international best practice. NSW should use that laboratory to ensure AI used by NSW and in NSW is safe and subject to ongoing monitoring and assurance.**

NSW should be commended for factoring in secondary or cumulative harms in its consideration of risk.[37] NSW is right that the harms of AI systems might not be felt by the person who receives the product of the service, and that trust is a relevant consideration. The framework calls on the assessor to think deeply about everyone who might be impacted, well beyond the obvious end user.

NSW should update its guidance regarding secondary harms to include the implications of engaging any particular AI developer – including the reputational benefit for that developer and the implications of it receiving further funding. NSW should seek to only deal with AI developers with the strongest possible commitments to AI ethics and AI safety – including demonstrated investments in

[36] "Generative AI: Implications for Trust and Governance". Infocomm Media Development Authority, Singapore & Aicadium. (2023).
https://www.gov.uk/government/news/initial-100-million-for-expert-taskforce-to-help-uk-build-and-adopt-next-generation-of-safe-ai
https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf
[37] NSW Aritifcal intelligence assurance framework, Page 24.

and commitments of computing resources to longer-term AI safety considerations. We are likely to see a proliferation of AI developers with a range of risk tolerances for the potential harms of their products and a range of commitments to longer-term research into addressing those harms. Commercial incentives that reward risk-taking AI developers could contribute to catastrophic second-order harms, and avoiding this should be an explicit consideration.

We recommend:

- **NSW should update its assessment procedure of secondary harms to include an assessment of the commitment of NSW's commercial partners towards longer-term AI safety and AI ethics.** NSW should only do business with the most scrupulous AI developers. We need a world where AI labs are appropriately cautious Supporting risk-taking by unethical labs could have catastrophic second-order consequences.

The role of NSW in the Federation

NSW begun its AI-journey in a positive direction, and the above recommendations are intended to ensure it remains a leader. Despite NSW's success, early signs from other jurisdictions were less positive. The Commonwealth Department of Industry, Science and Resources ongoing "Supporting Responsible AI" consultation has shown worrying signs.

One issue with the Commonwealth's approach so far is that it neglects the need to prepare for risks that might still be a few years in the future. Hundreds of AI experts are raising the alarm about risks from highly capable AI, including through the Statement on AI Risk and the call for a Pause on Giant AI experiments. In a survey of experts in the field, 48% of respondents gave at least a 10% chance of an extremely bad outcome from AI.[38] One in five Australians believe AI presents a risk of human extinction in the next 20 years, and 57% believe AI will create more problems than it solves - including job losses, but also highlighting the need for regulation, that AI can be misused, and the unknown consequences from

[38] Stein-Perlman, Z., Weinstein-Raun, B., Grace, K., (2022). *2022 Expert Survey on Progress in AI.* AI Impacts. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai

developing and deploying some AI systems.[39] Polling from the US shows that most Americans think AI will achieve greater than human levels of intelligence and think that it should be subject to strong regulation, akin to medical devices. A majority of Americans support blunt instruments like a pause on AI research, and like Australians, 1 in 5 Americans think AI could be an existential risk to humanity.[40]

Despite these serious and widespread concerns, the Commonwealth's Discussion Paper did not mention these issues. Australia's Chief Scientist observes that these kinds of risks are at least two or five years away, "difficult to forecast", and so does not engage with the topic.[41] CSIRO acknowledges the possibility that AI is an existential threat and due diligence is necessary, but minimises the concern because the threat is not "imminent".[42]

At a recent public town hall event, one of the Commonwealth's key advisors compared thinking about these big risks as being similar to the Wright Brothers thinking about how to regulate a Mars colony. While this is a humorous image, being confident that advanced AIs are in the order of 150 years into the future is inconsistent with the weight of expertise and inconsistent with NSW's acknowledgement of the pace of change and transformative nature of AI.

Similarly, a senior public servant at the same event analogised AI to a kitchen knife, arguing that it might be dangerous in narrow circumstances, but is an essential everyday item that likely does not require specific regulation. This view is wrong and dangerous. Products like "undress AIs" are disanalogous to a kitchen knife, serve no productive place in our society, and are immediately harmful. There's also no pending "step change" in kitchen knife capability. We don't need to worry that kitchen knives already have dangerous dual-use capabilities that could lead to widespread or catastrophic harm, and there aren't current experiments showing concerning progress towards autonomous kitchen knives unaligned with the intentions of their designers. Each of these issues is discussed in more detail above.

---

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

[40] Elsey et al. (2023). *US public onion of AI Policy and risk.* Rethink Priorites. https://rethinkpriorities.org/publications/us-public-opinion-of-ai-policy-and-risk

[41] Australia's Chief Scientist. (2023). *Rapid Response to Information Report: Generative AI*. Pages 1 and 10. https://www.chiefscientist.gov.au/GenerativeAI

[42] CSIRO. Whittle et al. (2023). *Hype or fear: the AI debate examined.* https://www.csiro.au/en/news/All/Articles/2023/June/AI-debate-examined

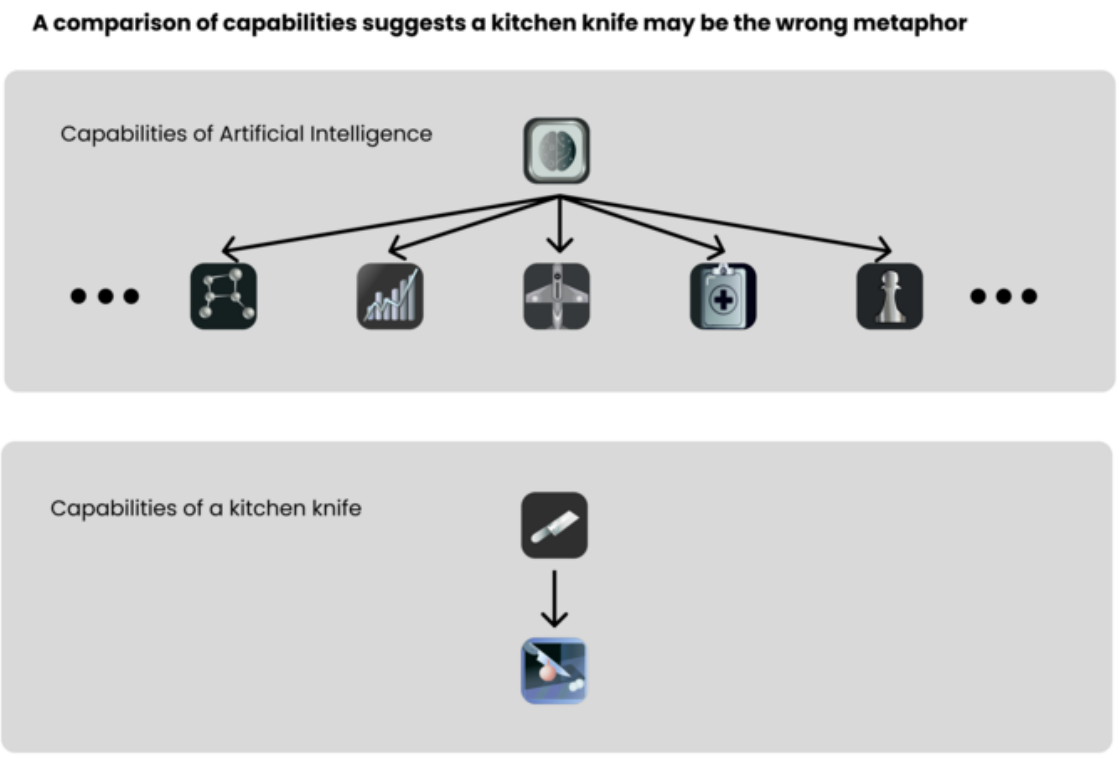

Figure 1. A comparison of capabilities suggests a kitchen knife may be the wrong metaphor

Overall, if the Commonwealth adopts a "gold rush" perspective that focuses only on maximising economic output and worries only about risks that are already occurring, Australia could be left in a dangerous position.

The UK's 1 November 2023 AI Safety Summit may provide an opportunity for the Commonwealth to change course. The Summit includes session on many of the topics discussed in this document, including:

- Risks to Global Safety from Frontier AI Misuse
- Risks from Unpredictable Advances in Frontier AI Capability
- Risks from Loss of Control over Frontier AI
- What should National Policymakers do in relation to the risk and opportunities of AI?

In light of this national trend, NSW may need to guide the thinking of other jurisdictions through forums like National Cabinet, and perhaps act as a "wicketkeeper" to protect the residents of NSW and citizens of Australia from the

potential harms of AI if other jurisdictions are slow to act. NSW has shown that it understands the pace of change and the transformative nature of AI – and it should maintain that forward-leaning approach.