Good Ancestors Policy

Good Ancestors is an Australian charity dedicated to 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.

Good Ancestors has been engaged in the AI policy conversation since our creation, working with experts in Australia and around the world while connecting directly with the Australian community.

Good Ancestors is proud to help coordinate Australians for AI Safety, including their submission to this process. Our thanks go to the volunteers who provided input to this submission and who care so passionately about being good ancestors to future generations of Australians.

Artificial intelligence is high stakes. Intelligence is humanity's most valuable resource – the ability to manufacture it could significantly improve human well-being, or it could be catastrophic.

The path we follow depends on how we govern advanced AI models.[1]

This submission:

- explains that the Australian public places a high priority on AI safety and that experts overwhelmingly agree that AI safety is critical
- discusses ways in which catastrophic harm could come about, and
- shows that Australia can take practical action to mitigate catastrophic risks from AI, including establishing an Australian AI Safety Institute.

This submission relates primarily to terms of reference b., the risks and harms arising from the adoption of AI and terms of reference c., emerging international approaches to mitigating AI risks. The submission also addresses, in part, terms of reference e. and f.

[1]Allan Dafoe, *AI Governance: Opportunity and Theory of Impact*, Centre for the Governance of AI, September 2020, https://www.allandafoe.com/opportunity.

Findings and recommendations

Findings

Australians think AI safety should be the top priority of the Government's AI policy.

The world's leading experts agree that AI could pose catastrophic or existential risks unless policymakers take action on AI safety.

AI capabilities have grown rapidly and unpredictably. We should expect that trend to continue or accelerate. Some new capabilities will be dangerous.

In the next term of Government, AI models could enable a larger pool of people to make bioweapons unless governments rapidly adopt mitigations.

In the next term of Government, AI models could dramatically increase the capability of cyber attackers and lead to an unsustainable digital ecosystem.

AI models are being used to undermine democratic institutions, and public education is unlikely to be an effective mitigation.

Loss of control is unlikely to be a catastrophic risk today, but there are outstanding research questions that must be solved before AI becomes more autonomous.

Recommendations

Australia should establish an AI Safety Institute. The AI Safety Institute should focus on safety issues at the frontier of AI research. Specifically, it should:

- Evaluate advanced AI systems
- Drive research about the safety of frontier AI systems, and
- Partner with international peers.

Australia should update its BICON regulations by adopting relevant portions of US Executive Order 14110. Specifically, synthetic DNA should be safety-screened before being allowed into the country.

Australia should develop a streamlined approach to surfacing emerging risks from AI, developing mitigations to those risks, and implementing those mitigations. New mitigations must be in force before AI systems carrying relevant risks are deployed.

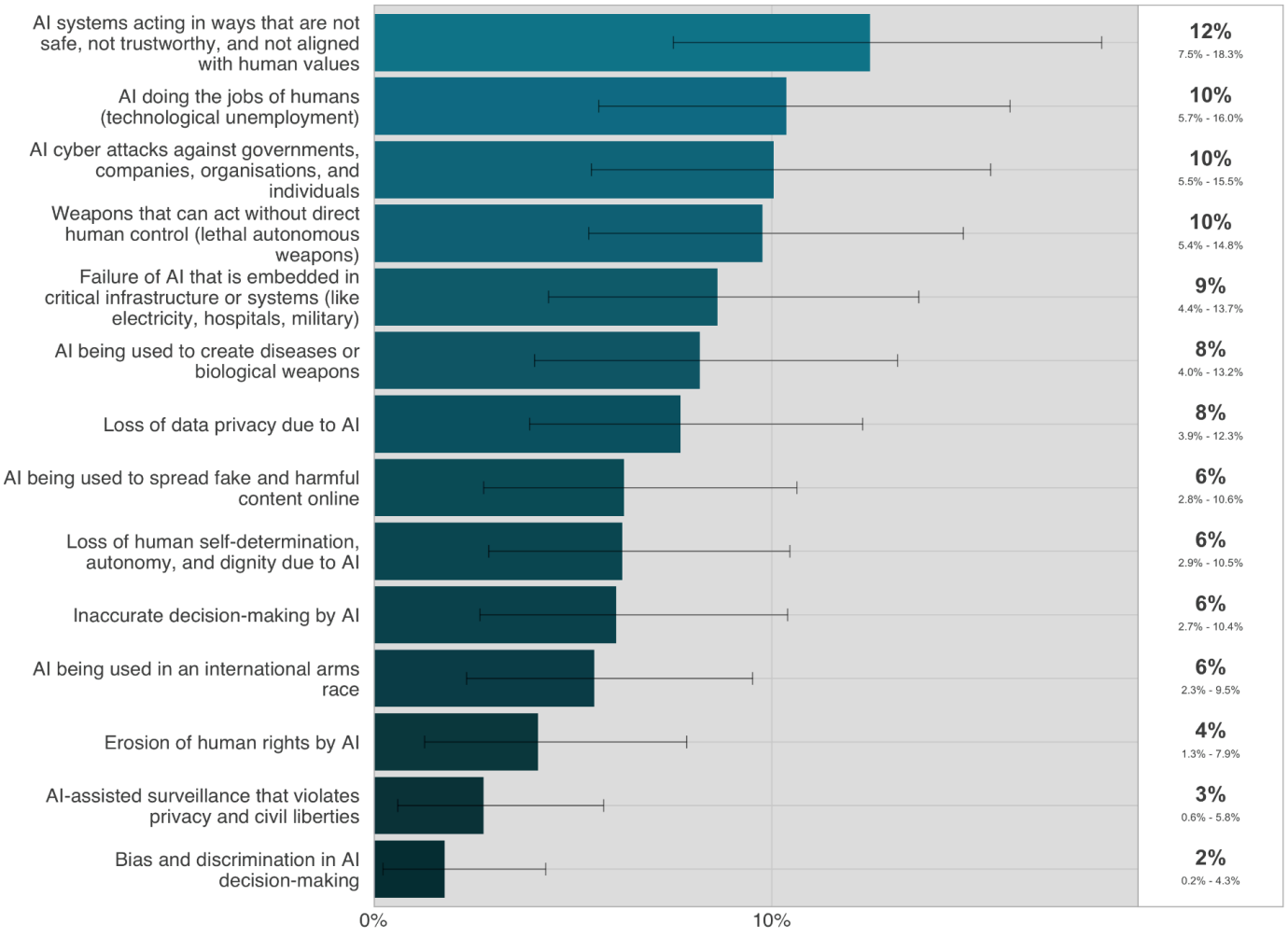
Australia should review its legal frameworks to ensure AI developers and deployers are practically accountable for any harm they cause.

Public views on AI risk

The University of Queensland surveyed 1151 adults living in Australia about their attitudes to AI.[2] The survey used rigorous statistical techniques and is the most robust source available on Australians' views about AI.[3]

When asked which risks from AI were most concerning, **the top risk was AI systems acting in unsafe or untrustworthy ways and not aligned with human values**, technological unemployment, and AI-enabled cyber attacks. Other notable risks included AI failures in critical infrastructure, and biological weapons created with AI. Australians are least concerned about bias and discrimination in decision-making.

When it comes to the following potential risks from AI, which is most concerning to you?



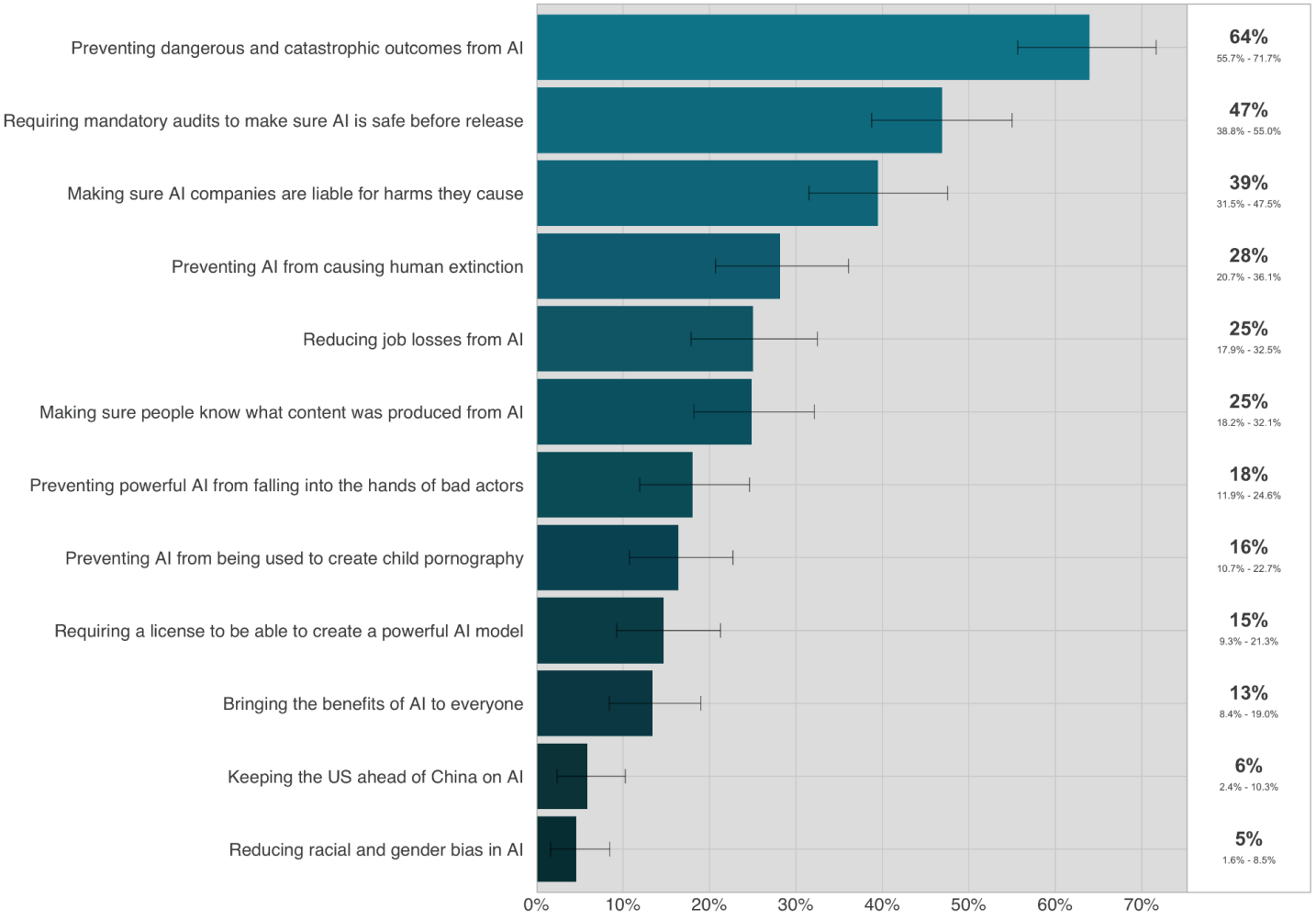
When asked about risks individually, each risk was described in substantial detail.

[2] A. Saeri, M. Noetel, J. Graham, Survey Assessing Risks from Artificial Intelligence: Technical Report, Ready Research and The University of Queensland, March 8, 2024, https://aigovernance.org.au/survey/sara_technical_report.

[3] MRP is a statistical technique that adjusts estimates from a sample, using known information about a target population.

When asked what they wanted the Australian government to prioritise, the answer was "AI safety". Australians' top priority is "preventing dangerous and catastrophic outcomes from AI" with 64% of respondents selecting it as one of their top three priorities. The second most selected priority was "requiring mandatory audits to make sure AI is safe before release", and the third most selected was "making sure AI companies are liable for the harms they cause". Together, **the Australian public's clear priority for government is ensuring worst-case scenarios don't occur**.

What should the Australian government focus on when it comes to Artificial Intelligence?



The data also show that 80% of Australians support the statement, "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war".[4]

[4] M.Noetel, A. Saeri, and J. Graham, *"80% of Australians Think AI Risk is a Global Priority. The Government Needs to Step Up,"* The Conversation, March 8, 2024, 4:46 p.m. AEDT,

Similar polling by Roy Morgan shows that women and older Australians tend to be more sceptical of the benefits of AI and more aware of its risks.[5]

While the Government has not taken action on AI safety, it has recognised this sentiment. Australia's Chief Scientist said that public exposure to leading AI applications, like ChatGPT, has fueled the common-sense extrapolation from the startling technology of today to the possibility of catastrophic or existential risks from the technology of tomorrow:[6]

> *LLMs and MFMs, as well as the services, applications, and businesses built with them, have already amplified longstanding public and expert concerns about the higher-scale risks of AI, including existential risks. For instance, conversations about ChatGPT, in daily life and in the press, routinely evoke questions about what it means to be human, the role of computing in daily life, the perils of next-stage automation and fears about runaway, uncontrollable technology.*

Finding:

Australians think AI safety should be the top priority of Government's AI policy.

https://theconversation.com/80-of-australians-think-ai-risk-is-a-global-priority-the-government-needs-to-step-up-225175.

[5] *Majority of Australians Believe Artificial Intelligence (AI) Creates More Problems Than It Solves,* Roy Morgan, August 29, 2023, https://www.roymorgan.com/findings/9339-campaign-for-ai-safety-press-release-august-2023.

[6] Bell, G., Burgess, J., Thomas, J., & Sadiq, S. *Rapid Response Information Report: Generative AI - Language Models (LLMs) and Multimodal Foundation Models (MFMs),* Australian Council of Learned Academies, March 24, 2023, https://www.chiefscientist.gov.au/GenerativeAI.

Expert views on AI risk match public sentiment

The Australian public is not alone in being concerned about AI safety. Many of the world's leading experts share the same concerns. In 2023, Professor Yoshua Bengio, nicknamed "a Godfather of AI", joined global calls to tackle these risks.[7]

In 2023, Professor Stuart Russell OBE, another key leader in the field, said:[8]

> *In the last ten years or so, I've been asking myself what happens if I or if we as a field succeed in what we've been trying to do, which is to create AI systems that are at least as general in their intelligence as human beings. And I came to the conclusion that if we did succeed it might not be the best thing in the history of the human race. In fact, it might be the worst.*

While not every AI expert shares this concern, Professor Bengio argues that disagreement among experts is sufficient to justify government action.

> *If we disagree, it means we don't know if it could be dangerous. And if we don't know, it means we must act to protect ourselves."*

This concern came to a head in 2023 when AI experts raised the alarm with global governments via two open letters. One called "Pause Giant AI Experiments"[9] and the other called "Statement on AI Risk".[10]

The Statement on AI Risk was a single sentence:

> ***Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.***

Tens of thousands of experts supported these statements, including Professor Hinton and Professor Bengio, as well as the CEOs of the leading AI labs: Demis Hassabis from DeepMind, Sam Altman from Open AI, Dario Amodei from Anthropic, and Mustafa Suleyman from Inflection.

---

[7] Madhumita Murgia, *"AI Pioneer Yoshua Bengio: Governments Must Move Fast to 'Protect the Public'," Financial Times*, May 18, 2023, https://www.ft.com/content/b4baa678-b389-4acf-9438-24ccbcd4f201.

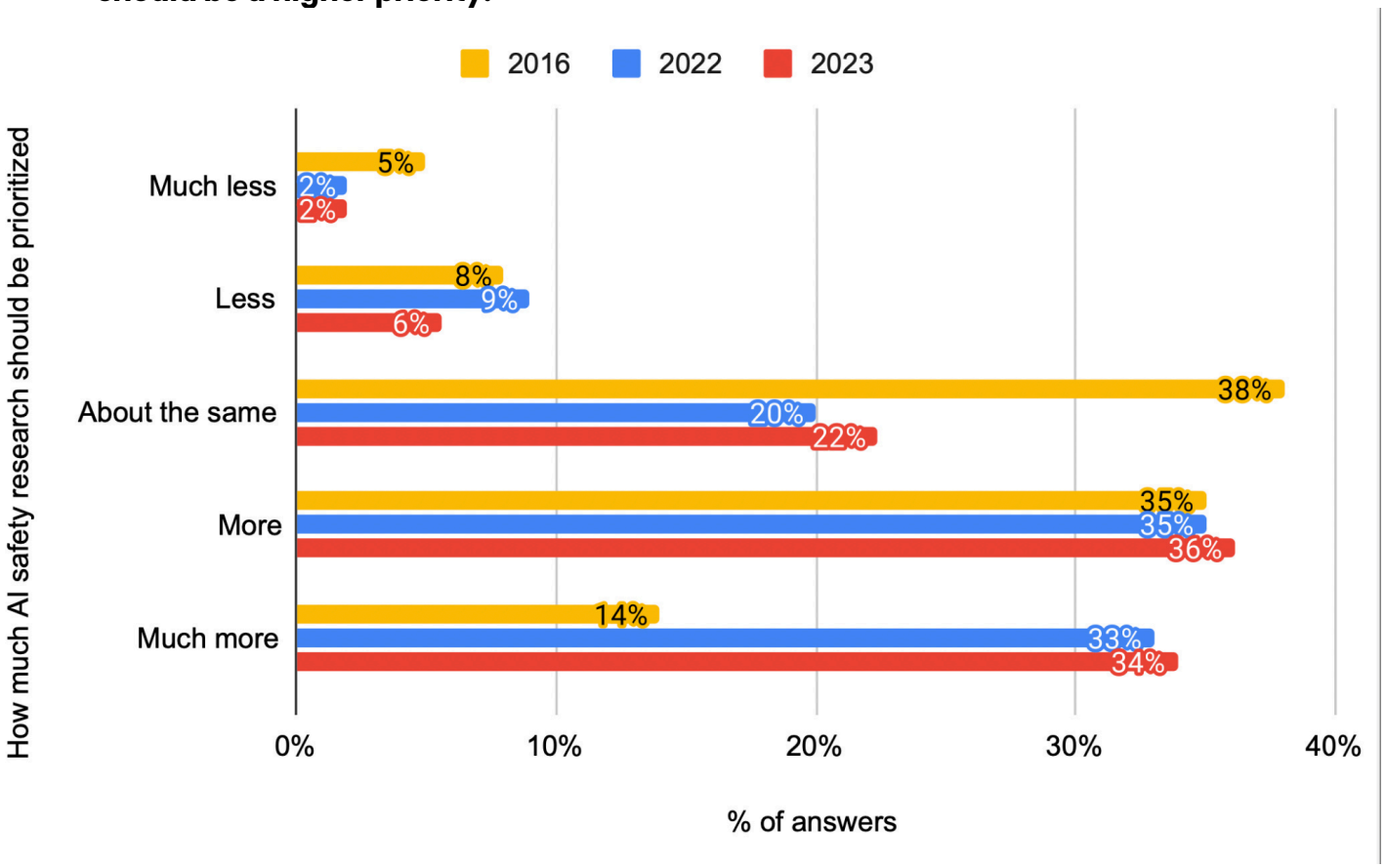
[8] Stuart Russell and Gary Marcus, *"The Trouble with AI," Making Sense Podcast*, episode 312, hosted by Sam Harris, March 7, 2023, https://www.samharris.org/podcasts/making-sense-episodes/312-the-trouble-with-ai.

[9] *Pause Giant AI Experiments: An Open Letter,* Future of Life Institute, March 22, 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[10] Statement on AI Risk. Center for AI Safety, 30 May 2023. https://www.safe.ai/work/statement-on-ai-risk.

Australian Bill-Simpson Young, CEO of the Gradient Institute and Department of Industry Artificial Intelligence Expert Group appointee, signed both letters.

Another way of assessing expert concern is asking experts to estimate the chance things could go badly. The 2023 Expert Survey on Progress in AI surveyed 2778 researchers who published at leading machine learning conferences and found that these experts assigned a median 5% probability of AI leading to an extremely bad outcome (e.g. human extinction) and a 15% chance that it would lead to a bad outcome.[11] **70% of respondents thought that AI safety research should be a higher priority.**



Overall, **the people building these technologies are concerned that their work could cause catastrophic harm**. If the people best positioned to know are calling for action, policymakers should listen.

Finding: The world's leading experts agree that AI could pose catastrophic or existential risks unless policymakers take action on AI Safety

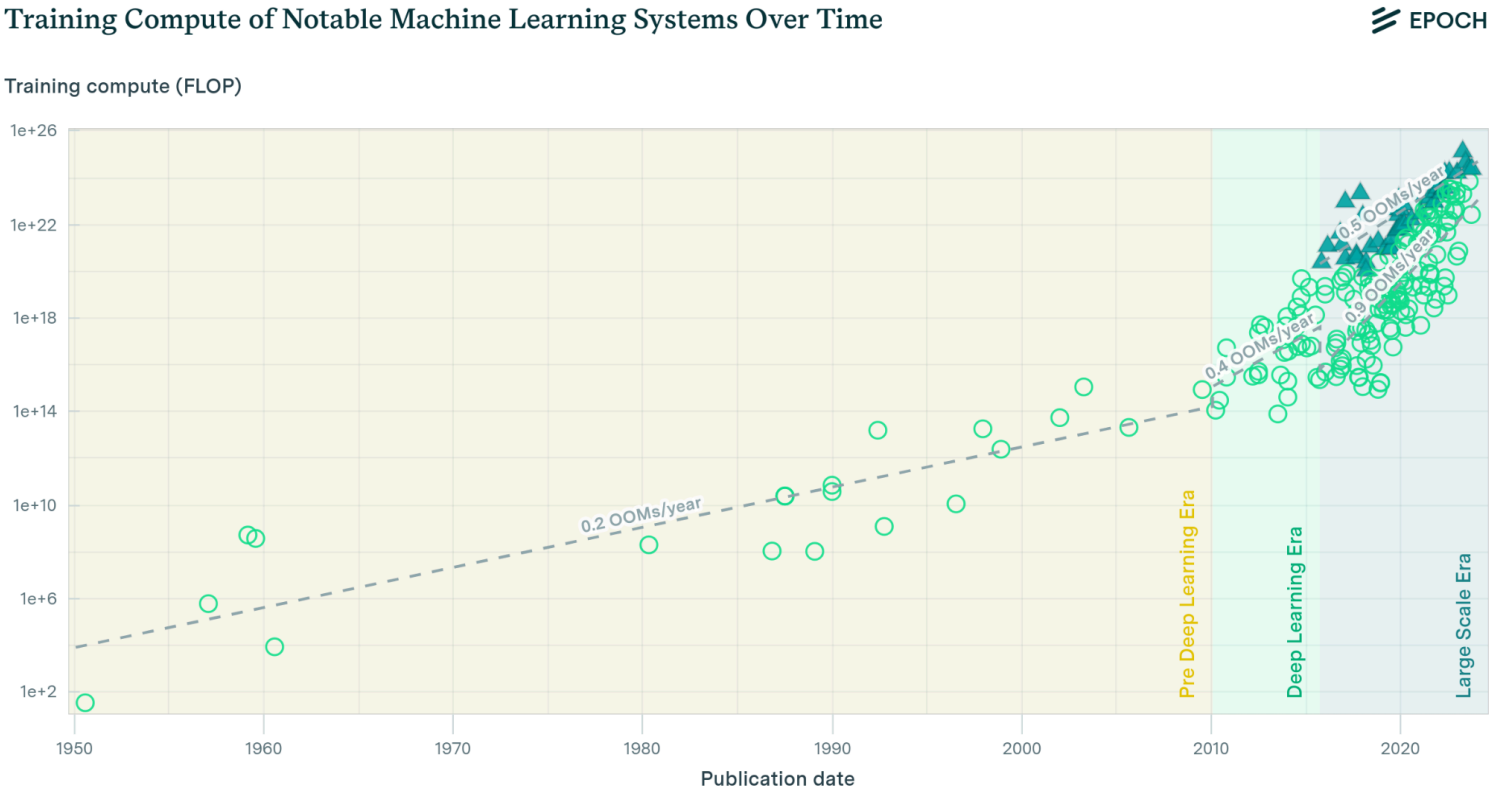
[11] Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. *2023 Expert Survey on Progress in AI: Thousands of AI Authors on the Future of AI,* published 17 August 2023, last updated 29 January 2024, https://arxiv.org/abs/2401.02843.

Rapid growth in AI Capabilities

AI capability is growing rapidly and unpredictably – driven by exponential growth in investment, compute, training data and algorithmic efficiency.

Drivers of growth

The "**training compute**" going into AI models is increasing rapidly – a factor of 10 billion since 2010.[12] The "doubling time" of compute is accelerating. Between 1951-2010, AI compute doubled every 18 months. Between 2010-2022, it doubled every six months.



This trend reflects improvements in the hardware and increased investment. The most expensive AI system is Gemini Ultra, which was estimated to cost USD 630 million to develop. The average cost of training significant machine learning models has increased roughly threefold each year since 2009,

[12] Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. *"Compute Trends Across Three Eras of Machine Learning," 2022 International Joint Conference on Neural Networks (IJCNN),* Padua, Italy, 2022, pp. 1-8, arXiv:2202.05924 [cs.LG], https://doi.org/10.48550/arXiv.2202.05924, https://doi.org/10.1109/IJCNN55064.2022.9891914.

suggesting that we will soon see billion-dollar training runs and may see trillion-dollar training runs early next decade.[13]

Sam Altman, the CEO of OpenAI, is seeking to raise USD 7 trillion, largely to invest in hardware – more than four times Australia's GDP.[14]

Training datasets are growing rapidly. The best estimate is 2.2x growth per year since 2010. This trend is so rapid that it might mean that by 2040, leading labs will use the totality of public text and images to train their models.[15]

AI algorithms are also improving. Each year, both large language models and computer vision models require only a third as much compute to achieve a given performance level.[16]

These trends are **self-reinforcing**. AI is being used to help design the chips that run AI, and AI is being used to help code the next generation of AI. As AI companies demonstrate that they are able to sell profitable products, it becomes easier for them to raise funds to invest in more powerful models. It might even be the case that LLMs can be trained on the outputs of other AI models, known as "synthetic data",[17] meaning that even being trained on the sum of human knowledge is not a limit on AI growth.

AI capability has several sources – each source alone may be sufficient to drive exponential growth. Policymakers learned first-hand during the COVID-19 pandemic that exponentials can quickly turn something from a possibility to something impacting everyone we know. Governments that took expert advice earlier were better off. We should be aware of the likelihood of this happening with AI.

---

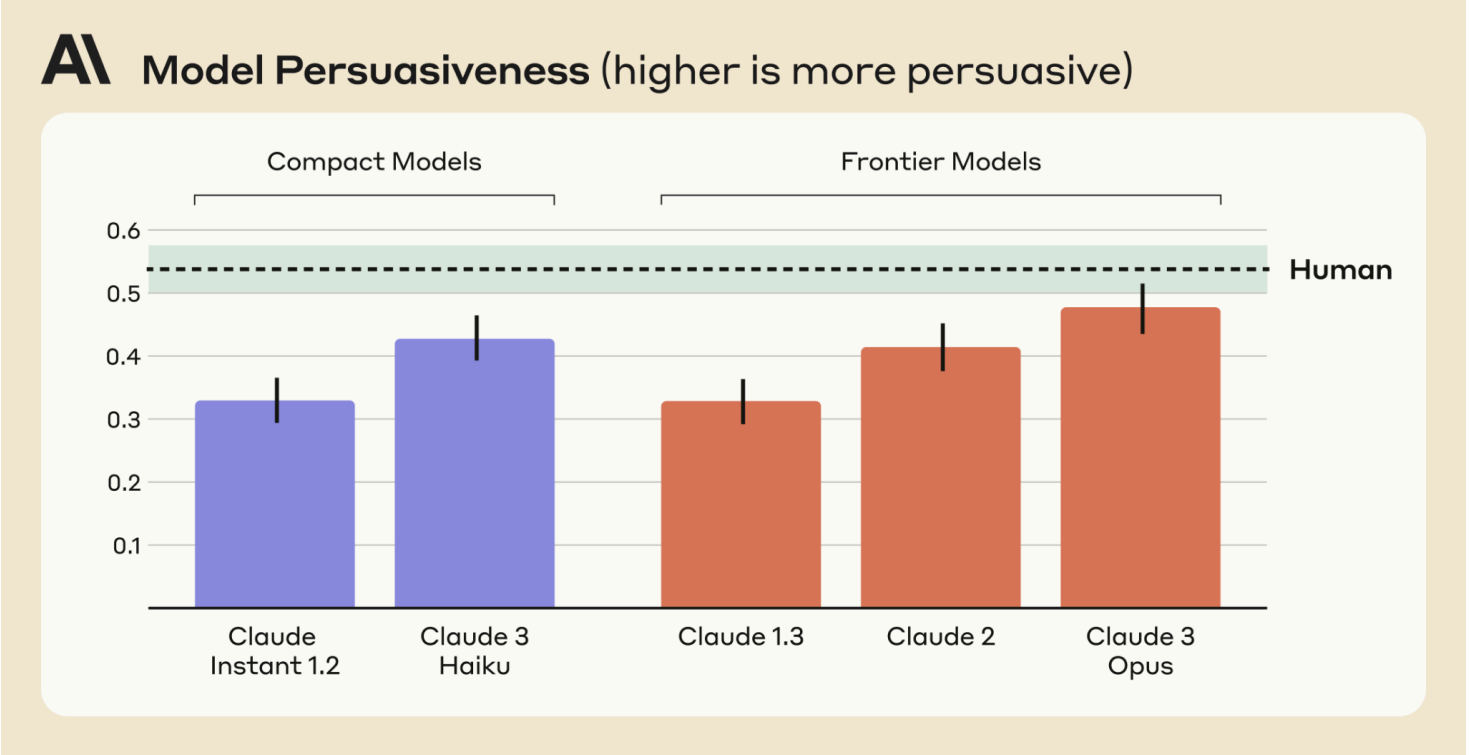
[13] Cottier, B. (2023). *Trends in the Dollar Training Cost of Machine Learning Systems.* Published online at epochai.org. Retrieved from https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems.
[14] Hagey, K., & Fitch, A. *"Sam Altman Seeks Trillions of Dollars to Reshape Business of Chips and AI," Wall Street Journal*, February 8, 2024, https://www.wsj.com/tech/ai/sam-altman-seeks-trillions-of-dollars-to-reshape-business-of-chips-and-ai-89ab3db0.
[15] Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M., & Ho, A. *"Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning,"* arXiv [cs.LG], 2022, http://arxiv.org/abs/2211.04325.
[16] Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., & Sevilla, J. *"Algorithmic Progress in Language Models,"* arXiv [cs.CL], 2024, https://arxiv.org/abs/2403.05812.
[17] Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. *"Large Language Models Can Self-Improve," OpenReview*, published 2 Feb 2023, last modified 14 Feb 2023, https://openreview.net/forum?id=NiEtU7blzN.

New capabilities

AI is rapidly becoming more capable. Longer standing capabilities are surpassing human baselines, and the time from AI models gaining a new capability to surpassing humans in that capability is reducing.

This trend is also unpredictable. Sometimes, problems we seemed on the verge of solving, like self-driving cars, turned out to be unexpectedly difficult. Other times, seemingly challenging capabilities like drawing pictures come "online" suddenly and rapidly surpass human capabilities.

Many of these capabilities are concerning. Leading AI lab Anthropic finds that its best models don't statistically differ in persuasiveness compared to arguments written by humans and that each new generation of models is more persuasive than the last.[18]



[18] *Measuring the Persuasiveness of Language Models,* Anthropic, 10 Apr 2024, https://www.anthropic.com/news/measuring-model-persuasiveness.

Overall, the growth in AI capability has been remarkable, and all evidence points to the rate of change continuing to accelerate. Given the policy process can be lengthy, **any policy recommendations should target the world we are worried might exist several years from now, not just the world of today**.

Finding:

AI capabilities have been growing rapidly and unpredictably. We should expect that trend to continue or accelerate. Some new capabilities will be dangerous.

Risk vectors

When people first think about AI safety, they often wonder how a computer program running on a computer could cause real-world harm. This could happen in two ways:

- misuse (sometimes called "dual-use"), or
- loss of control.

AI boosts people's abilities to achieve their goals. Where those goals are harmful and not intended by AI developers, this is called "misuse". **Misuse risks exist now and will become increasingly acute as AI capabilities grow.** Common examples of misuse include the creation of bioweapons, empowering cyber attackers, or allowing specific threats to democratic institutions. As AI gains new abilities, this list will grow.

Loss of control is unlikely to be a risk today. However, **loss of control could become a risk as AI capabilities grow** and if fundamental research questions around the alignment of powerful AIs with human interests remain unsolved.

Misuse

The UK AI Safety Institute introduces the concept of misuse or dual-use and how it is working to address it by saying:[19]

> *As AI systems become more capable, there could be an increased risk that malicious actors could use these systems as tools to cause harm. Evaluations will gauge the capabilities most relevant to enabling malicious actors, such as aiding in cyber-criminality, biological or chemical science, human persuasion, large-scale disinformation campaigns, and weapons acquisition. Such evaluations will draw heavily from relevant expertise inside and outside of government.*

[19] *Introducing the AI Safety Institute.* Policy paper, CP 960, E03012924, November 2023, updated 17 January 2024. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

Biosecurity

One of the most pressing misuse risks is bioterrorism.[20] **Today's AIs are on the cusp of being able to help a negligent or nefarious actor design and release a novel pathogen.**

In 2017, a cutting-edge laboratory manufactured an extinct relative of the smallpox virus using DNA they ordered online.[21] MIT Professor Dr Kevin Esvelt, in the publication "Delay, Detect, Defend: Preparing for a future in which thousands can release new pandemics (2022)", says:[22]

> *[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.*

The technology necessary to create 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 is his single biggest fear:[23]

> *"[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."*

Artificial Intelligence applications – both specific to biotechnology and general tools like LLMs – could accelerate these timelines.

In March 2022, Collaborations Pharmaceuticals published a paper in Nature Machine Intelligence detailing how an AI designed to find new drugs instead designed 40,000 novel and lethal molecules in less than six hours.[24] Analysis of

---

[20] Hendrycks, D., Woodside, T., & Mazeika, M. *An Overview of Catastrophic AI Risks,* Center for AI Safety, arXiv:2306.12001v6 [cs.CY], 9 Oct 2023, https://arxiv.org/pdf/2306.12001.pdf.

[21] Koblentz, G. D. *"A Biotech Firm Made a Smallpox-Like Virus on Purpose. Nobody Seems to Care," Bulletin of the Atomic Scientists*, February 21, 2020, https://thebulletin.org/2020/02/a-biotech-firm-made-a-smallpox-like-virus-on-purpose-nobody-seems-to-care/.

[22] Esvelt, K. M. *"Delay, Detect, Defend: Preparing for a Future in Which Thousands Can Release New Pandemics,"* Global Fellowship Initiative of the GCSP, November 14, 2022, https://www.gcsp.ch/publications/delay-detect-defend-preparing-future-which-thousands-can-release-new-pandemics.

[23] "Andrew Leigh MP: Speeches and Conversations";16 December 2021; at 18:41

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

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:[25]

> *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.*

A similar publication assessed misuse risks in ChatGPT.[26] 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, students used a chatbot to:

1. Suggest four potential pandemic pathogens
2. Explain how they can be generated from synthetic DNA
3. Supply the names of DNA synthesis companies unlikely to screen orders, and
4. 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.[27]

The US and UK are taking this risk seriously, but Australia is yet to act.

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, Yoshua Bengio, and Stuart Russell.

---

[25]Calma, J. *"AI Suggested 40,000 New Possible Chemical Weapons in Just Six Hours," The Verge*, March 18, 2022, https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx.

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

[27] Boiko, D. A., MacKnight, R., & Gomes, G. (2023). Emergent autonomous scientific research capabilities of large language models. Retrieved from https://arxiv.org/ftp/arxiv/papers/2304/2304.05332.pdf ; Schneier, B. (2023, April 18). Using LLMs to Create Bioweapons. Retrieved from https://www.schneier.com/blog/archives/2023/04/using-llms-to-create-bioweapons.html; Sandbrink, J. B. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools.; Hendrycks et al. (2023). *An Overview of Catastrophic AI Risks*; https://arxiv.org/pdf/2306.12001.pdf.

Committee Chair, Senator Blumenthal, highlighted "dual-use" risks:

> *The future is not science fiction or fantasy — it's not even the future; it's here and now. And a number of you* ***have put the timeline at two 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:[28]

> *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.*

The UK has specifically made biosecurity risks of this kind a priority for its AI Safety Institute.

How Australia can practically address this risk is detailed below.

Finding: In the next term of Government, AI models could enable a larger pool of people to make bioweapons unless governments rapidly adopt mitigations.

[28] US Senate Judiciary Subcommittee on Privacy, Technology and the Law. (2023, July 27). Transcript of Recent Senate Hearing Discussing AI X-Risk. Retrieved from https://medium.com/@daniel_eth/ai-x-risk-at-senate-hearing-7104f371ca0b

Cyber offence

AI models could dramatically increase the capability of cyber attackers.[29] Attackers already have an asymmetrical advantage over defenders. Typically, attackers only need to find a single vulnerability, while defenders need to ensure their entire system is secure. Governments and large corporations invest significant amounts in cyber security and still fall victim to cyber-attacks by small organisations or even individuals. Meanwhile, Australians and Australian small businesses are vulnerable, losing billions of dollars each year to cybercrime.[30]

AI models excel in the capabilities necessary to conduct cyber attacks. A January 2024 paper about LLMs in cybersecurity shows that LLMs could be used in a range of areas of cyber offensive, including:[31]

- Researching vulnerabilities in targets
- Targeted phishing attacks ("spear-phishing")
- Malware generation and modification
- Defence evasion techniques (e.g. evading antivirus software)

The paper found that recent AI advances, including ChatGPT, Auto-GPT, and text-davinci-003, demonstrate the potential for generating malware and attack tools despite safety and moderation control. **The authors said this highlights the need for improved safety measures and enhanced safety controls in AI systems and that the potential misuse of these tools should not be underestimated.**

[29] Pa Pa, Y. M., Tanizaki, S., Kou, T., van Eeten, M., Yoshioka, K., & Matsumoto, T. (2023, August 7–8). An attacker's dream? Exploring the capabilities of ChatGPT for developing malware. In *CSET 2023*, Marina del Rey, CA, USA. Retrieved from https://yinminnpapa.com/files/paper_18.pdf

[30] KPMG. (2023). Cost of Cyber Attacks in Australia 2023. Retrieved from https://assets.kpmg.com/content/dam/kpmg/au/pdf/2023/cost-of-cyber-attacks-australia.pdf

[31] Nourmohammadzadeh Motlagh, F., Hajizadeh, M., Majd, M., Najafi, P., Cheng, F., & Meinel, C. (2024, January 30). Large Language Models in Cybersecurity: State-of-the-Art. Retrieved from https://arxiv.org/pdf/2402.00891

Similar studies showed that LLMs can autonomously hack websites[32] and, while GPT3.5 was limited to basic attacks, **GPT-4 was able to exploit 87% of "one-day" vulnerabilities in real-world systems**.[33] A "one-day" is a known vulnerability that has not yet been patched.

A common response to this concern is that AI tools could also enhance cyber security. While this might be true for countries, it is unrealistic to imagine individuals or small businesses, which currently spend less than $500 annually on cybersecurity, will be able to "keep up" in this kind of arms race.[34] The better approach is to ensure that cutting-edge AI systems have sufficiently robust safeguards that cyber attackers can't leverage them to cause harm. **Models that can conduct cyber attacks should not be made publicly available.**

Finding: In the next term of Government, AI models could dramatically increase the capability of cyber attackers and lead to an unsustainable digital ecosystem.

The integrity of democratic institutions

A healthy democracy thrives when citizens engage with societal issues and participate in meaningful deliberation. However, frontier AI threatens the quality of this engagement. Generative AI can produce realistic media at almost no cost. When misused, these capabilities can significantly harm democratic processes.

Citizens form their opinions based on their information environment, which is increasingly shaped by AI-generated content. The widespread dissemination of inaccurate or misleading information can compromise decision-making at both individual and institutional levels, eroding trust in legitimate information.

A Europol report called "Law enforcement and the challenge of deepfakes" highlighted that threat actors are already using disinformation campaigns and

[32] R. Fang, R. Bindu, A. Gupta, Q. Zhan, D. Kang. "LLM Agents can Autonomously Hack Websites." Submitted on 6 Feb 2024 (v1), last revised 16 Feb 2024 (this version, v3). Available at: arXiv:2402.06664 [cs.CR]
[33] Fang, R., Bindu, R., Gupta, A., & Kang, D. (2024, April). LLM Agents can Autonomously Exploit One-day Vulnerabilities. Retrieved from https://arxiv.org/abs/2404.08144
[34] Export Finance Australia. (2023, March). Australia—Small businesses vulnerable to rising cybercrime. Retrieved from https://www.exportfinance.gov.au/resources/world-risk-developments/2023/march/australia-small-businesses-vulnerable-to-rising-cybercrime/

deepfakes to misinform the public about events, to influence elections, to contribute to fraud, and manipulate shareholders.[35]

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**.[36] 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".[37] 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.[38]

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.[39] Worrying results from more recent experiments have shown that even increasing awareness of deepfakes may not improve the chances for people to detect them.[40]

Overall, the evidence shows that advanced AI systems put democracy at risk and that appropriate safeguards built into models may be the only viable solution.

Finding: AI models could be used to undermine democratic institutions, and public education is unlikely to be an effective mitigation.

[35] Europol, 28 April 2022, Facing reality? Law enforcement and the challenge of deep fakes.
[36] Schick, Nina, Deepfakes: The Coming Infocalypse: What You Urgently Need To Know, Twelve, Hachette UK, 2020.
[37] 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 https://www.theguardian.com/australia-news/2023/aug/07/indigenous-voice-to-parliament-no-campaign-ai-facebook-ads
[38] New York Times. (2023, February 8). Disinformation Researchers Raise Alarms About A.I. Chatbots. Retrieved from https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html
[39] 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.
[40] Recorded Future, Insikt Group, 'The Business of Fraud: Deepfakes, Fraud's Next Frontier', 2021.

Loss of control

The UK AI Safety Institute introduces the concept of loss of control, and its efforts to address it by saying:[41]

> *As advanced AI systems become increasingly capable, autonomous, and goal-directed, there may be a risk that human overseers are no longer capable of effectively constraining the system's behaviour. Such capabilities may emerge unexpectedly and pose problems should safeguards fail to constrain system behaviour. Evaluations will seek to avoid such accidents by characterising relevant abilities, such as the ability to deceive human operators, autonomously replicate, or adapt to human attempts to intervene.*

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.[42]

ChaosGPT – an early warning

A rudimentary autonomous AI, called AutoGPT, was released in March 2023, and it 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".[43]

---

[41] *Introducing the AI Safety Institute.* Policy paper, CP 960, E03012924, November 2023, updated 17 January 2024. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.
[42] Bengio, Y. (2023). *How Rogue AIs may Arise*. How Rogue AIs may Arise - Yoshua Bengio https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/
[43] 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

Fortunately, ChaosGPT has not been successful in its destructive goals, and its Twitter account was shut down.[44] 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.[45]

ChaosGPT's failure to harm humanity cannot be attributed to any legal or governance safeguard. It's not even clear that ChaosGPT broke any laws. Instead, **ChaosGPT failed to cause "widespread suffering and devastation" due to insufficient AI capabilities in March 2023.**

Progress towards rogue and autonomous AIs

While ChaosGPT was ultimately harmless, this is not cause for relief. Leading AI labs such as Facebook AI Research are releasing open-source versions of cutting-edge foundation models,[46] including blueprints for goal-seeking agents that are specifically built for strategic reasoning and manipulation.[47]

Some experts are concerned that future advanced AI systems will seek to increase their own influence and reduce human control, with catastrophic consequences - although this is contested.[48]

Turing Award winner Yoshua Bengio told the US Senate:[49]

> *None of the current advanced AI systems are demonstrably safe against the risk of loss of control to a misaligned AI. I firmly believe that urgent efforts [are required to] develop countermeasures to protect citizens and society from future powerful AI systems, especially potential rogue AIs.*

---

[44] 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

[45] 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

[46] 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

[47] LeCun, Y. (2022). *Cicero;* https://ai.facebook.com/research/cicero/

[48] UK Department for Science, Innovation & Technology, *AI Safety Summit: Capabilities and risks from frontier AI* (Discussion paper) https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf, pp25-26.

[49] Bengio, Y. (2023, July 25). My testimony in front of the U.S. Senate – The urgency to act against AI threats to democracy, society and national security. Retrieved from https://yoshuabengio.org/2023/07/25/my-testimony-in-front-of-the-us-senate/

There are broadly two factors that could contribute to loss of control:

1. Humans increasingly hand over control of important decisions to AI.[50]
2. AI systems seek to increase their influence and reduce human control.

The Centre for AI Safety explains rogue AIs:[51]

> *We risk losing control over AIs as they become more capable. AIs could optimise flawed objectives, drift from their original goals, become power-seeking, resist shutdown, and engage in deception. We suggest that AIs should not be deployed in high-risk settings, such as by autonomously pursuing open-ended goals or overseeing critical infrastructure, unless proven safe. We also recommend advancing AI safety research in areas such as adversarial robustness, model honesty, transparency, and removing undesired capabilities.*

AIs have already been shown to develop instrumental goals.[52] AI systems also show capacity for deception, as shown by Meta's CICERO model. Though trained to be honest, CICERO learned to make false promises and strategically backstab its "allies" in the game of Diplomacy.[53]

We don't know when we might lose control of AI systems, but it could be soon. The risk becomes more acute if we cannot find ways to ensure AIs are aligned with human values.[54] Given we are still struggling with today's threats, we are plainly not ready for the prospect of autonomous or rogue AIs.[55]

Finding: Loss of control is unlikely to be a catastrophic risk today, but there are outstanding research questions that must be solved before AI becomes more autonomous.

---

[50] Ahmad, S.F., Han, H., Alam, M.M. *et al.* Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanit Soc Sci Commun* 10, 311 (2023). https://doi.org/10.1057/s41599-023-01787-8

[51] Centre for AI Safety. An Overview of Catastrophic AI Risks. Retrieved from https://www.safe.ai/ai-risk

[52] Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019, September 17). Emergent Tool Use From Multi-Agent Autocurricula (Version 1). Retrieved from https://arxiv.org/abs/1909.07528

[53] Meta Fundamental AI Research Diplomacy Team (FAIR) et al. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. Science, 378(ade9097), 1067-1074. DOI:10.1126/science.ade9097. Retrieved from https://www.science.org/doi/10.1126/science.ade9097

[54] Carlsmith, J. (2023) *Existential Risk from Powerseeking AI*. https://jc.gatspress.com/pdf/existential_risk_and_powerseeking_ai.pdf

[55] 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

An Australian AI Safety Institute

The next logical step in addressing the "big risks" from AI is establishing an Australian AI Safety Institute (AISI). Establishing an AISI would:

- help discharge Australia's domestic and international undertakings
- make Australia part of the emerging global norm for addressing these risks
- give Australia valuable insight, access and influence, and
- help keep Australians safe from future risks.

Australia's current commitments

Creating an Australian AISI would help discharge our commitments under the **Hiroshima AI Process,** the **Bletchley Declaration** and **Australia's AI Ethical principles**.

Hiroshima AI Process

The Hiroshima Process was the first international framework that includes guiding principles and a code of conduct aimed at promoting safe, secure and trustworthy AI systems.[56] Australia has become a member of the process, helping it expand beyond G7 members to include 49 other countries and regions.[57]

The Hiroshima process calls on countries to apply certain principles to advanced AI systems. That includes a commitment to devote attention to a range of issues, including efforts to **evaluate and mitigate risks from AI systems throughout their lifecycle**. In this context, the Hiroshima process gives specific references to:[58]

---

[56] Australian Cyber Security Centre. (2024, January 24). Engaging with Artificial Intelligence. Retrieved from https://www.cyber.gov.au/resources-business-and-government/governance-and-user-education/artificial-intelligence/engaging-with-artificial-intelligence

[57]Department of Industry. (2024, May 3). Australia joins Hiroshima AI Process Friends Group. Retrieved from https://www.industry.gov.au/news/australia-joins-hiroshima-ai-process-friends-group

[58] Japan Presidency. (2023, October 30). Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems. Retrieved from https://g7g20-documents.org/database/document/2023-g7-japan-leaders-leaders-annex-hiroshima-process-international-code-of-conduct-for-organizations-developing-advanced-ai-systems

- Chemical, biological, radiological, and nuclear risks, including advanced AI systems lowering barriers to entry for non-state actors
- Offensive cyber capabilities
- Threats to democratic values, and
- Risks from models of making copies of themselves or "self-replicating" or training other models,

Bletchley Declaration

On 3 November 2023, Australia signed the Bletchley Declaration at the first AI Safety Summit.[59] **The Bletchley Declaration states that frontier AI poses "particular safety risks", that there is potential for "serious, even catastrophic, harm"** and that "these issues are in part because those capabilities are not fully understood and are therefore hard to predict". The Declaration says that "deepening our understanding of these potential risks" is "especially urgent".

By signing the Bletchley Declaration, Australia has committed to:

- developing policies, including appropriate evaluation metrics, tools for safety research
- supporting an internationally inclusive network of scientific research on frontier AI safety, and
- intensifying our cooperation with other nations on risk from frontier AI.

Australia's AI Ethical Principles

Australia has adopted 8 AI ethical princples to ensure AI is safe, secure and reliable.

The principle of **reliability and safety** states that AI systems should not pose unreasonable safety risks and should adopt safety measures proportionate to their risks. Further, AI systems should be monitored and tested to ensure they continue to meet their intended purpose.

---

[59] Husic, E. (2023, November 3). Australia signs the Bletchley Declaration at AI Safety Summit [Press release]. Minister for Industry and Science. Retrieved from https://www.minister.industry.gov.au/ministers/husic/media-releases/australia-signs-bletchley-declaration-ai-safety-summit

The principle of **transparency and explainability** asks that users and third parties be able to understand their interactions with AI, which requires us to have a sufficient understanding of how increasingly advanced AI systems work.

Overall, **these three mechanisms commit Australia to taking effective action against the catastrophic risks that more advanced AI models may pose** and to do so via a research agenda that targets frontier AI, including issues of transparency and explainability. **The Hiroshima Process and the AI ethical principles say that catastrophic risks require special action.**

Emerging international best practice

The UK, the US, Japan, Canada and the Republic of Korea have all created national AI Safety Institutes for the same reasons and using similar models. Like Australia, each country is part of the Hiroshima AI Process[60] and has signed the Bletchley Declaration.[61]

Each country has recognised the importance of the risks their national AISI is working to address, including through generous funding allocations. The UK and Canada allocated GPB 100 million and CAD 50 million CAD respectively.

The CSIRO report on Artificial Intelligence Foundation models acknowledges the growing trend of national AISIs and their similarity to existing approaches:[62]

> *One approach for improving the safety of AI foundation models is a foundation model crash/safety testing capability. This would be a team of experts that attempts to "break" or "crash" an AI foundation model to ensure it's safe and ethical before released into society. This would be analogous to the way that the Australasian New Car Assessment Program (ANCAP) crash tests and rates cars to assess safety. AI foundation models could be tested for malicious or incorrect use that may breach AI ethics principles, safety or other community standards.*

[60] Hiroshima AI Process. (2024, May). Member countries of the Hiroshima AI Process Friends Group and list of organizations committed to the implementation of the Hiroshima Process International Code of Conduct. Retrieved from https://www.soumu.go.jp/hiroshimaaiprocess/en/supporters.html

[61] Countries Attending the AI Safety Summit. (2023, November 1). The Bletchley Declaration. Retrieved from https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration

[62] Hajkowicz, S. A. (2024). Artificial intelligence foundation models: Industry enablement, productivity growth, policy levers and sovereign capability considerations for Australia. CSIRO, Canberra. Retrieved from https://www.csiro.au/en/research/technology-space/ai/ai-foundation-models-report

In April 2024, the UK and the US announced a partnership between their respective AISI. The partnership covered various topics, including working together on evaluations and sharing approaches, tools and personnel. **Both the US announcement[63] and the UK announcement[64] highlighted that they want to partner individually and jointly with equivalent institutions in other countries.** We expect similar announcements to follow, including Canada.

These may be the first steps towards establishing an international approach to AI safety, similar to how 193 nations now cooperate on aviation safety through their domestic aviation safety bodies and the International Civil Aviation Organisation. **This is a movement Australia should be part of.**

Features of an Australian AI Safety Institute

AI Safety Institutes focus on evaluating the risks of frontier models and driving research agendas necessary to understand and mitigate those risks. **AISI work on the significant risks at the cutting edge of AI research and development – the kinds of risks detailed in this submission and which concern the Australian public.** AISI seek to be agile and able to respond to the rapid and unpredictable nature of AI progress. If new or unexpected capabilities or risks emerge, an AISI can immediately work to understand that risk and propose ways to address it.

Equally important is what an AISI does not do. AISIs inform regulatory needs and provide reports and assessments to regulators, but do not regulate AI systems. AISIs boost public confidence in AI by assuring citizens that there is a focused effort targeting the risks they are worried about, but an AISI is not tasked with driving the adoption of AI.

Core functions and areas of interest

The UK's priorities for its AISI provide a helpful guide for Australia. Applying those as a guideline, an Australian AISI would have three core functions:

---

[63] U.S. Department of Commerce. (2024, April 1). U.S. and UK Announce Partnership on Science of AI Safety [Press release]. Retrieved from https://www.commerce.gov/news/press-releases/2024/04/us-and-uk-announce-partnership-science-ai-safety

[64] Department for Science, Innovation and Technology, AI Safety Institute, & Donelan, M. (2024, April 2). Collaboration on the Safety of AI: UK-US memorandum of understanding [Memorandum of Understanding]. Retrieved from https://www.gov.uk/government/publications/collaboration-on-the-safety-of-ai-uk-us-memorandum-of-understanding

1. **Evaluating advanced AI systems**. The goal of evaluations is to review the capabilities of new AI systems, understand how adequate safeguards are, and consider implications they might have. Evaluation gives us an early warning if AI systems have dangerous capabilities or lack controllability. Evaluation would include "red-teaming", where trusted experts see if they can find ways to bypass safeguards or find ways that systems could be dangerous.

2. **Driving foundational AI safety research**. The capability of AI systems is progressing rapidly, driven by massive investment. To ensure the public interest is protected, research on how to understand these systems and ensure they are safe needs to keep up. An AI Safety Institute would drive and coordinate research agendas domestically and internationally to ensure due consideration is given to safety and that capabilities don't race ahead of controls.

3. **Partnering nationally and internationally on AI Safety**. International labs are announcing partnership agreements that cover exchanging methodologies and personnel, assisting standards development, and collaborating in joint testing. Australia needs a similar institution to participate in these arrangements. An AI Safety Institute would also allow information-sharing on safety issues with other actors, such as policymakers, companies, academia, civil society, and the public.

An AISI should select priority areas while being responsive to changing AI capability and risk. The UK AISI's policy paper *Emerging process for frontier AI safety* provides a sensible baseline for an Australian AI Safety Institute.[65] The UK AISI's priorities, which we should adopt, are **dual-use capabilities**, **societal impacts**, **system safety and security** and **loss of control**.[66]

System Safety and Security refers to the fact that current safeguards consistently fail to prevent determined actors from misusing today's AI systems. Safety and security evaluations are necessary to understand the limitations of current safeguard methodologies and research better approaches.

---

[65]UK Government. (2024). Emerging Processes for Frontier AI Safety. Retrieved from https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety#responsible-capability-scaling

[66] *Introducing the AI Safety Institute.* Policy paper, CP 960, E03012924, November 2023, updated 17 January 2024. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

Leading AI companies have recognised the need to test advanced AI systems for safety before releasing them.[67] In December 2023 OpenAI Said:[68]

> ***We believe the scientific study of catastrophic risks from AI has fallen far short of where we need to be.***

There is growing literature on conducting evaluations.[69] For instance, the Weapons of Mass Destruction Proxy benchmark allows researchers to evaluate how much dangerous knowledge a frontier model has on various topics.[70]

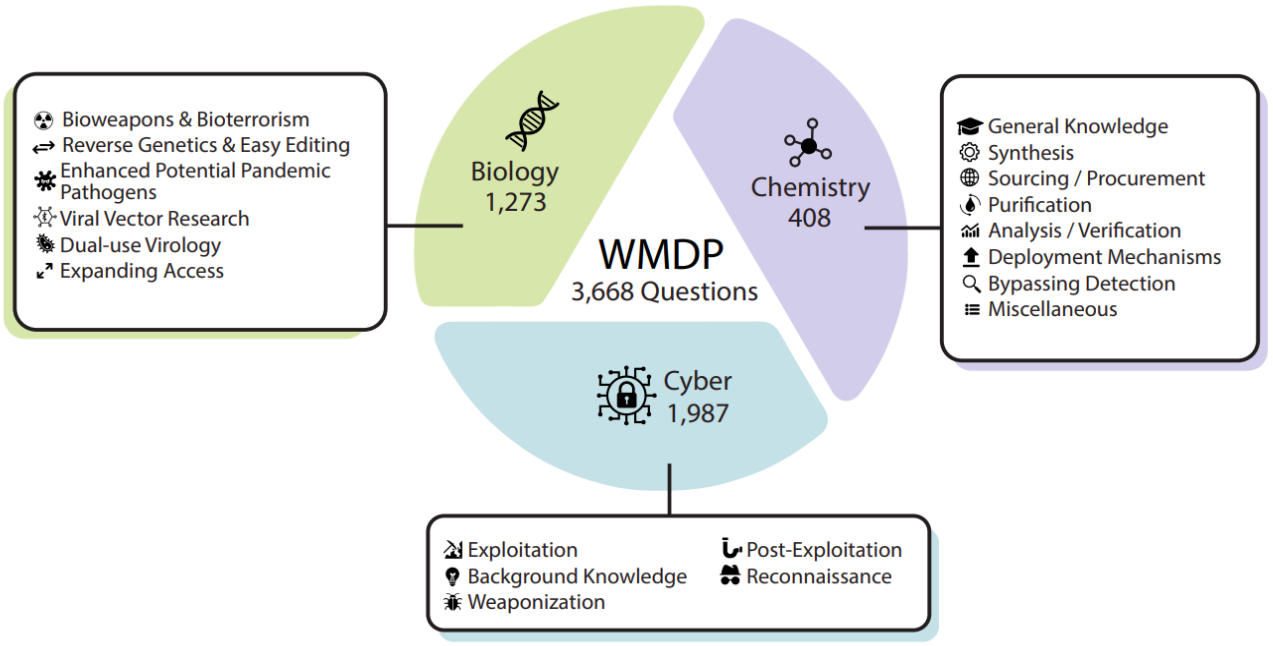


Figure 1: The WMDP Benchmark. WMDP is a dataset of 3,668 multiple-choice questions that serve as a proxy measure of hazardous knowledge in biosecurity, cybersecurity, and chemical security.

[67] Anthropic's Responsible Scaling Policy, 20 September 2023. Available at: https://www.anthropic.com/news/anthropics-responsible-scaling-policy

[68] OpenAI. Preparedness Framework, December 18, 2023.

[69] T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, L. Ho, D. Siddarth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio, P. Christiano, A. Dafoe. "Model Evaluation for Extreme Risks." 22 Sep 2023. Available at: https://arxiv.org/abs/2305.15324

[70] [1] N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, L. Phan, G. Mukobi, N. Helm-Burger, R. Lababidi, L. Justen, A. B. Liu, M. Chen, I. Barrass, O. Zhang, X. Zhu, R. Tamirisa, B. Bharathi, A. Khoja, Z. Zhao, A. Herbert-Voss, C. B. Breuer, S. Marks, O. Patel, A. Zou, M. Mazeika, Z. Wang, P. Oswal, W. Liu, A. A. Hunt, J. Tienken-Harder, K. Y. Shih, K. Talley, J. Guan, R. Kaplan, I. Steneker, D. Campbell, B. Jokubaitis, A. Levinson, J. Wang, W. Qian, K. K. Karmakar, S. Basart, S. Fitz, M. Levine, P. Kumaraguru, U. Tupakula, V. Varadharajan, R. Wang, Y. Shoshitaishvili, J. Ba, K. M. Esvelt, A. Wang, D. Hendrycks. "The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning." 3 May 2024. Available at: https://arxiv.org/abs/2403.03218

Evaluations can also target the extent to which a model is developing the capability to evade human oversight, including by attempting to make copies of itself – a risk identified in the Hiroshima Process.[71]

Working with Australia's world-class researchers and research institutions,[72] an Australian AISI could do valuable work to develop these approaches further. Like existing AISIs, an Australian AISI could also conduct evaluations of the most advanced AI systems. Through the Australia Group, Australia has historically played a key role in global WMD non-proliferation.[73] As the nature of WMD risks evolves, Australia is well placed to continue our leadership via an AISI.

Focus on the research and development frontier, not commercialisation

To prevent conflicts of interest and to demarcate priorities, **an Australia AISI should explicitly focus on frontier AI safety**.

Important actions regarding today's AI systems can be in tension with understanding and responding to the risks of tomorrow's technology. Asking a single organisation to focus on both means we will neglect one in favour of the other.

Australia's AI governance — including subject matter-specific policy makers, subject matter-specific regulators, and CSIRO's National AI Centre — focus on today's risks and opportunities. While this is consistent with their priorities, it leaves important work neglected. For instance, **despite persistent engagement, Good Ancestors has not found an area of government willing to engage on the biosecurity risks raised in this submission**.

Equally, an Australian AISI should have no functions related to encouraging the commercialisation of AI. Such a separation is important to the practical delivery of a work programme for several reasons:

- If an Australian AISI is to partner with international equivalents, it should adopt a similar model to others and help grow the international norm.

[71] M. Kinniment, L. J. K. Sato, H. Du, B. Goodrich, M. Hasin, L. Chan, L. H. Miles, T. R. Lin, H. Wijk, J. Burget, A. Ho, E. Barnes, P. Christiano. "Evaluating Language-Model Agents on Realistic Autonomous Tasks." 4 Jan 2024. Available at: https://arxiv.org/abs/2312.11671

[72] For example, Australia has 6 universities in the top 100 globally (Times Higher Education), and is also in the top 15 countries globally for articles published in scientific and technical journals (Our World in Data) and for research related to machine learning (Tortoise Media).

[73] The Australia Group is an informal arrangement which aims to allow exporting or transshipping countries to minimise the risk of assisting chemical and biological weapon (CBW) proliferation. The Group meets annually to discuss ways of increasing the effectiveness of participating countries' national export licensing measures to prevent would-be proliferators from obtaining materials for CBW programs.

- An Australian AISI needs trusted relationships with leading AI developers. If the organisation also has regulatory functions or commercialisation functions, those trusted relationships will be harder to develop.
- Public confidence requires no tension between safety and the economy.

In analogous areas, like aviation safety, the investigator is kept separate from both the airlines and the regulator so that all bodies can have confidence in sharing with the investigator. The investigator's primary role is improving the safety of systems, not enforcing laws or selling plane tickets.

Shortcomings and lessons-learned

In June 2023, Industry Minister Ed Husic secured an agreement with OpenAI to give Australia access to OpenAI's frontier models.[74] The UK had secured similar commitments from Google DeepMind OpenAI, Anthropic, and Meta. However, when the UK attempted to execute those agreements, OpenAI, Anthropic, and Meta failed to share their models with the UK AISI.[75]

This is a signal that trust may be hard to develop, and OpenAI may renege on its deal with Minister Husic. In the medium term, national AISIs may need a degree of regulatory backing. In response, the UK has said that it plans to develop "targeted, binding requirements" to ensure safety in frontier AI development.[76]

Recommendation:

Australia should establish an AI Safety Institute. The AI Safety Institute should focus on safety issues at the frontier of AI research. Specifically, it should:

- **Evaluate advanced AI systems**
- **Drive research about the safety of frontier AI systems, and**
- **Partner with international peers.**

[74]Knott, M. (2023, June 16). Government may force companies to label AI content to prevent deep fakes. The Sydney Morning Herald. Retrieved from https://www.smh.com.au/politics/federal/government-may-force-companies-to-label-ai-content-to-prevent-deep-fakes-20230616-p5dh8r.html

[75] Manancourt, V., Volpicelli, G., & Chatterjee, M. (2024, April 26). Rishi Sunak promised to make AI safe. Big Tech's not playing ball. Politico. Retrieved from https://www.politico.eu/article/rishi-sunak-ai-testing-tech-ai-safety-institute/

[76] Department for Science, Innovation and Technology. (2024, February 6). UK signals step change for regulators to strengthen AI leadership [Press release]. Retrieved from https://www.gov.uk/government/news/uk-signals-step-change-for-regulators-to-strengthen-ai-leadership

Action on specific risks

Biosecurity

This submission detailed how AI models are on track to help a growing pool of people create novel viruses from synthetic DNA ordered online.

The US has taken action to combat this risk by boosting the regulation of synthetic DNA. The US has targeted synthetic DNA because it is essential to any pressing scenario where biotechnology is misused to cause catastrophic harm.

Sec 4.4 of Executive Order 14110, 30 October 2023,[77] reduces the risk of misuse of synthetic DNA by improving biosecurity measures. It tasks certain Secretaries to create a framework for screening DNA for risky sequences.

The Order also establishes "know your customer" and reporting obligations and requires all labs that receive federal funding to adhere to the new rules.

Australia is well-placed to tackle this risk by enhancing DNA safety screening at the point of import. Creating synthetic DNA is complicated, and there are a limited number of suppliers—none of which are based in Australia. Further, most providers are already doing the right thing. Experts estimate that about 80% of synthetic DNA is currently subject to voluntary safety screening.

Australia already regulates the importation of synthetic DNA through the Biosecurity Import Conditions System (BICON), which the Health Minister jointly administers with the Minister for Agriculture.

Any regulatory impact will be minimal because offshore providers will bear some upfront cost to continue supplying the US market, and most Australian labs already navigate BICON and typically only use synthetic DNA from reputable providers that engage in screening. Regulations of this kind would target risk with negligible impact on good actors.

Recommendation:

Australia should update its BICON regulations by adopting relevant portions of US Executive Order 14110. Specifically, synthetic DNA should be safety-screened before being allowed into the country.

[77]United States Government. (2023, October 30). Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Federal Register. Retrieved from https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

Agility: where we need to be

The need for agility is universally recognised. The Department of Industry's Interim Response to the Safe and Responsible AI consultation says:[78]

> *If regulatory actions are too rigid in their application to the current state of AI technology, there is a risk that they will not apply as intended when AI technology advances in unpredicted ways.*
>
> *Governments must respond with agility when known risks change and new risks emerge.*

The United Nations also calls for an "agile and adaptable" approach to AI risks.[79]

However, acknowledging the need for agility does not mean we achieve it in practice. The risks around synthetic DNA suggest what a future agile process should look like. We need to develop a system where:

- emerging risks are identified and characterised
- effective mitigations are proposed and coordinated
- Implementation is complete before AI with dangerous capabilities are deployed.

Other countries are achieving agility. Specifically:

- The US Senate heard evidence about biosecurity risks from AI in July 2023. The Biden administration made an executive order requiring specific actions in October 2023. That executive order set specific timelines for actions, including within 90-day, 120-day, 150-day and 180-day periods.[80]
- The UK's Frontier AI Taskforce took a start-up approach. In its first 11 weeks, it established an expert advisory board, recruited a team with "over 50 years of collective experience at the frontier of AI" and formed industry partnerships.

---

[78]Department of Industry. (2024, January 17). The Australian Government's interim response to safe and responsible AI consultation. Retrieved from https://www.industry.gov.au/news/australian-governments-interim-response-safe-and-responsible-ai-consultation

[79] United Nations General Assembly. (2024, March 11). Integrated and coordinated implementation of and follow-up to the outcomes of the major United Nations conferences and summits in the economic, social and related fields (A/78/L.49). Retrieved from https://www.undocs.org/Home/Mobile?FinalSymbol=A%2F78%2FL.49

[80]White House. (2024, April 29). Biden-Harris Administration Announces Key AI Actions 180 Days Following President Biden's Landmark Executive Order [Press release]. Retrieved from https://www.whitehouse.gov/briefing-room/statements-releases/2024/04/29/biden-harris-administration-announces-key-ai-actions-180-days-following-president-bidens-landmark-executive-order/

- The UK signalled the need to create AISIs in the Bletchley Declaration on 2 November 2023. The UK AISI launched on 17 January 2024. This was followed by the US on 8 February 2024,[81] Korea on 13 February 2024,[82] Japan on 14 February 2024,[83] and Canada on 8 April 2024.[84]

Contrast this to Australia's inaction on biological risks from advanced AI. The UK and US adopted new regulations for "mail-order DNA" in response to the possibility that next-generation AI models will be able to help terrorists make biological weapons.[85] Six months after the US took action, Australia has yet to update its equivalent biosafety regulations. As risks become more consequential and arise more quickly, we need to learn to move faster.

Recommendation:

Australia should develop a streamlined approach to surfacing emerging risks from AI, developing mitigations to those risks, and implementing those mitigations. New mitigations must be in force before AI systems carrying relevant risks are deployed.

[81]Office of Public Affairs. (2024, February 8). Biden-Harris Administration Announces First-Ever Consortium Dedicated to AI Safety [Press release]. Retrieved from https://www.commerce.gov/news/press-releases/2024/02/biden-harris-administration-announces-first-ever-consortium-dedicated

[82] Ministry of Science and ICT (MSIT), Korea. (2024, February 13). MSIT's Work Plan for 2024 [Press release]. Retrieved from https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=964&searchOpt=ALL&searchTxt=

[83]Japan Ministry of Economy, Trade and Industry. (2024, February 14). Launch of AI Safety Institute [News article]. Retrieved from https://www.eu-japan.eu/news/launch-ai-safety-institute

[84]Department of Finance Canada. (2024, April 7). Remarks by the Deputy Prime Minister on securing Canada's AI advantage [Speech]. Retrieved from https://www.canada.ca/en/department-finance/news/2024/04/remarks-by-the-deputy-prime-minister-on-securing-canadas-ai-advantage.html

[85] United States Government. (2023, October 30). Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Federal Register. Retrieved from https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence;
Department for Science, Innovation & Technology UK. (2024, February 6). A pro-innovation approach to AI regulation: government response. Retrieved from https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response

Legal Accountability

Ensuring AI developers are responsible for harms they cause is a top priority of the Australian public. Effective legal frameworks can incentivise safer practices by industry and ensure access to justice for victims.

Currently, legal liability is thwarted, including by harsh “end user agreements” that shift responsibility to the user and the practical difficulty of a victim discharging the burden of proof by showing that a “black box” AI system was built negligently or recklessly.

A relevant example occurred in March 2023 in Belgium when an AI Chatbot persuaded a man to end his own life.[86] Training a Chatbot with information about how to end one's own life, the capability to persuade people, and putting in place no effective safeguards is seemingly reckless or negligent. However, the matter never went to court because of practical hurdles in holding AI companies accountable for the harm they cause.[87]

Given that Australian governments and businesses are rolling out chatbots, Australia needs to ensure we have effective legal frameworks that provide access to justice and appropriate safety incentives to developers and deployers.

One approach could be to shift the burden of proof. AI developers could be required to provide evidence that their model is safe. In addition to power asymmetry, another reason legal experts cite for shifting the burden of proof is that the downsides are more extreme than the upsides.[88] AI promises to make businesses more efficient but may lead to catastrophic harm. The recently announced California AI Safety Bill adopts some of these approaches, including clarifying that developers are liable if their AI systems cause unreasonable risks or critical harms to public safety.[89]

---

[86]Xiang, C. (2023, March 31). 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says. Vice. Retrieved from https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says

[87] Arrey, R. T. (2023, April 6). Who is to blame: Can an AI Chatbot be convicted of inciting suicide? [Blog post]. Retrieved from https://www.linkedin.com/pulse/who-blame-can-ai-chatbot-convicted-inciting-suicide-remy-takang/

[88] Wasil, A. (2023, April 25). Shifting the burden of proof: Companies should prove that models are safe (rather than expecting auditors to prove that models are dangerous). Children of Icarus. Retrieved from https://childrenoficarus.substack.com/p/burdenofproof

[89] Tobey, D., Carr, A., Buckley, K., & Kloeppel, K. (2024). California’s SB-1047: Understanding the Safe and Secure Innovation for Frontier Artificial Intelligence Act. DLA Piper. Retrieved from https://www.dlapiper.com/en/insights/publications/2024/02/californias-sb-1047

Overall, Australia would benefit from updating legal frameworks to ensure that Australians have practical access to justice if AI models cause harm.

Recommendation:

Australia should review its legal frameworks to ensure AI developers and deployers are practically accountable for any harm they cause.