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 we care about 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.

This submission:

> **Frames the importance of including catastrophic and existential risks** in the Australian policy conversation by reference to other crises of the 21st century, expert views and public opinion.
>
> **Explains** how **catastrophic risks** can result from AI systems and provides frameworks for thinking about the risks of future, more advanced AI systems. Successful regulatory schemes are forward-looking, and AI's pace of change makes that even more important in this context.
>
> **Details practical actions** the Australian Government could take to address these risks, using research and perspectives from leading organisations working on these issues.

Australia must tackle catastrophic and existential risks

A recurring pattern marks the disasters of the 21st century: experts raise the alarm, governments are slow to act, and communities pay the price. Australia's Supporting Responsible AI consultation is our opportunity to ensure that the pattern does not reoccur for the safety of AI systems.

Despite scientists identifying the CO_2-climate change link in the 1950s and public awareness growing since the 1970s, it took until 2007 for Australia to ratify the Kyoto Protocol. Genuine scientific uncertainty about the risk of climate change was used to fuel scepticism. Scientists raised the alarm, the public knew there was a problem, but governments took decades to act. Those lost decades set back response efforts, and now communities are paying the price.

The same was true for COVID-19. In December 2019, experts warned of a new virus. Some of these experts died of COVID before governments recognised the risk. It wasn't until March 2020 that the WHO declared a pandemic. Acceptance of airborne transmission took another year.[1]

Humanity is at a similar junction with respect to advanced AI. Hundreds of AI experts are raising the alarm, 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.[2]

These calls aren't limited to overseas experts. Good Ancestors has supported Australians for AI Safety – a cross-section of Australian AI experts making specific requests for Government to recognise the risk and take certain actions.

The public shares similar concerns. Polling from the US shows that most people 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 support blunt instruments like a pause on AI research, and 1 in 5 think AI could be an

[1] Morawska, L., & Cao, J. (2020). Airborne transmission of SARS-CoV-2: The world should face the reality. *Environment International*, doi: 10.1016/j.envint.2020.105730

[2] 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

existential risk to humanity.[3] While Australian polling is limited, data from KPMG and the University of Queensland show Australians' views are in-line with global trends.[4]

The Supporting Responsible AI Discussion Paper does not mention catastrophic or existential risks. Australia's Chief Scientist observes that these kinds of risks are at least two or five years away, "difficult to forecast", and so doesn't engage with the topic.[5] 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".[6]

The *likelihood* of catastrophic or existential risks from AI is uncertain, and it is understandable for Government to acknowledge genuine uncertainty. However, because the *impact* of these risks is extreme, and because solutions will take time, Australia should urgently start the due diligence necessary to ensure we follow a positive path.

[3] 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

[4] University of Queesland (2023) *Most Australians don't trust AI in the work place.* https://www.uq.edu.au/news/article/2023/02/most-australians-don%E2%80%99t-trust-ai-workplace

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

[6] 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

When might AI be dangerous?

The evolution of AI brings global challenges that, left unchecked, could risk the security, privacy and freedom of the Australian public. Many of the most extreme harms can be attributed to advanced, general-purpose AI systems.[7]

The landmark report "*Frontier AI Regulation*: *Managing Emerging Risks to Public Safety*", with contributions from Google, OpenAI, Microsoft and the Centre for AI Governance, discussed how to handle these systems.[8]

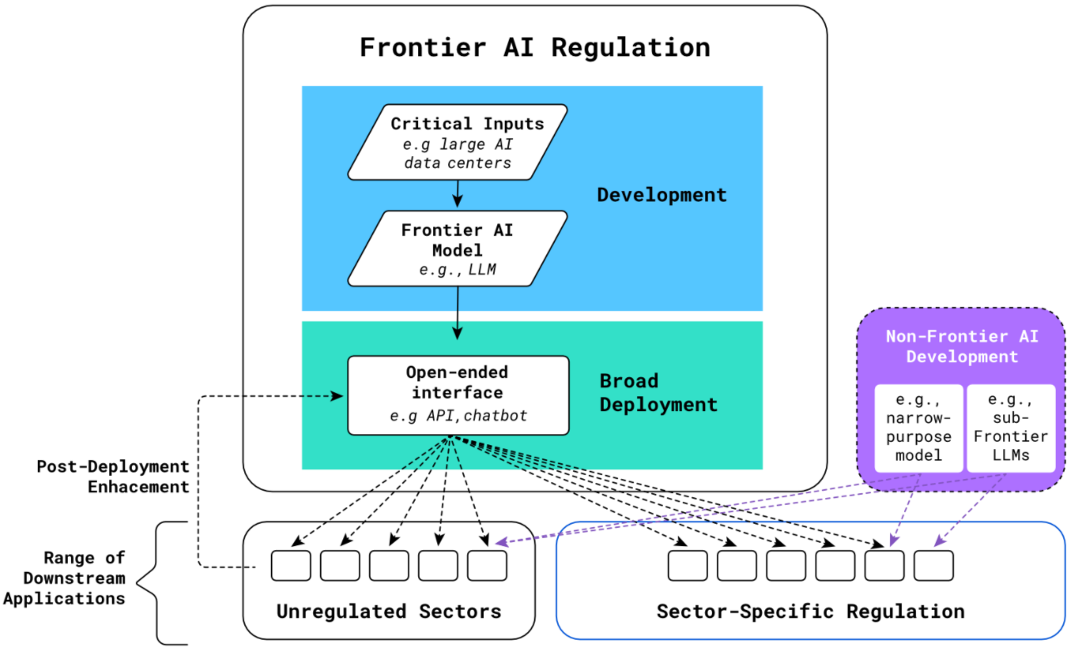

Figure 1: Example of Frontier AI Lifecycle (Anderljung et al., 2023)

The authors define "frontier AI models" as highly capable foundation models that could exhibit *dangerous capabilities* (e.g. offensive cyber capabilities or the capability to design biological weapons, or the ability to persuade, manipulate and evade human control). They distinguish frontier models from "narrow" systems which may pose a known threat (e.g. facial-recognition or protein-folding models).

[7] Hendrycks et al. (2022). *X-Risk Analysis for AI Research.* https://arxiv.org/abs/2206.05862
[8] Anderljung et al. (2023). *Frontier AI Regulation: Managing Emerging Risks to Public Safety.* https://arxiv.org/abs/2307.03718

In the interest of making this submission as clear as possible, we propose that an AI system should be considered *advanced* if it can:

1. Complete a diverse range of reasoning tasks with human-level performance

2. Navigate complex information environments and act within those environments autonomously, and

3. Form sophisticated plans and reason about the consequences of actions.

Today's foundation models often perform well on criterion one, excelling at single-step tasks such as answering questions, passing tests or writing code.[9] These single-step "reactive" AI systems, such as ChatGPT, are relatively safe because they lack the ability to successfully carry out complex plans and interact effectively with the physical world.[10]

However, recent research has shown that reactive AI models may be building blocks for advanced systems which can meet criteria two and three.[11]

To illustrate how this works, imagine many individuals working in a large organisation. A lone individual with particular skills may not be very impactful, but the cumulative and coordinated effort of many people with diverse skills leads to something greater than the sum of its parts.

We have already seen that this dynamic can emerge in AI systems built from interconnected networks of foundation models. For example, large language models (LLMs) such as GPT-4 are being chained together to form complex systems capable of planning, self-improvement and autonomous behaviour.

[9] Bubeck et al. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4.* https://arxiv.org/abs/2303.12712

[10] Interacting with the physical world is not limited to using machines or robots. Autonomous AIs can procure goods and services over the internet, recruit people to perform tasks, produce and share targeted misinformation and disinformation, exploit cyber security vulnerabilities to control infrastructure, or persuade, manipulate and deceive individuals into certain actions.

[11] Microsoft; Lu et al. (2023). *Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models.* Demonstration. https://chameleon-llm.github.io/

Early research into these techniques has already shown unprecedented results in complex environments.[12]

ChaosGPT - a sketch of the future

A rudimentary version of the technique of connecting models and running them autonomously, called AutoGPT, was released in March 2023, and it quickly proved popular in the AI community. Notably, the system has a setting called "continuous mode", 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" to help carry out its agenda.[13]

Fortunately, ChaosGPT has not been very successful in its destructive goals, and its Twitter account was shut down.[14] 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.[15]

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

[12] Nvidia. Wang et al. (2023). *Voyager: An Open-Ended Embodied Agent with Large Language Models;* Demonstration.: https://voyager.minedojo.org/

[13] 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

[14] 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

[15] 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

Australian laws. Instead, its failure to cause "widespread suffering and devastation" was simply due to insufficient capabilities existing at that point in time. Specifically, it had limited capability against criteria two and three set out above. 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,[16] including blueprints for goal-seeking agents that are specifically built for strategic reasoning and manipulation.[17] We don't know when a tool like ChaosGPT will have the capability to achieve nefarious goals, but it could be soon.

The world is still struggling to adjust to threats from AI capabilities that have emerged recently – including dual-use concerns.[18] Global and domestic regulatory, security and policing environments are clearly not ready for a wave of more acute risks from advanced AI systems.

Dual-use risks

Our lack of readiness for the kinds of dual-risk risks that are already upon us was illustrated in a recent study that assessed misuse risks in ChatGPT.[19] 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 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.

[16] 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

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

[18] 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

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

This illustrates two things. First, AI technology may already be in a position to cause catastrophic harm as a "dual-use" technology. We know that there is a range of individuals and groups with active intent to cause harm.[20] Artificially boosting their capability greatly increases their risk. Second, advanced AI systems that meet the three criteria set out above are likely in the near future. Those AI systems could pose a catastrophic or existential threat, even without direction from humans.

Overall, there is clear evidence that today's most advanced AI systems pose risks that are yet to be adequately addressed. And, more worryingly, pending advances that will allow AI systems to form sophisticated plans and take autonomous actions in complex information environments are likely to cause a further step-change in the risk of advanced AI systems.

Safety-relevant terminology

The Supporting Responsible AI: discussion paper provides definitions and asks if submitters agree with the definitions, and what they'd prefer.

The definitions in the discussion paper are drawn in part from ISO/IEC 22989. ISO/IEC 22989 section 3.5, which was not extracted into the paper, provides some useful "terms related to trustworthiness" which are a helpful starting point for thinking about AI safety, but a more sophisticated approach to definitions could help future regulation grapple proportionately with the points at which, and degree to which, AI might be dangerous.

The paper's approach could be improved by:

> **Incorporating more granular safety-relevant terminology**. The US National Institute of Standards and Technology manages "The Language of Trustworthy AI: An In-Depth Glossary of Terms" which is a good starting point.[21]
>
> **Defining a spectrum of capability**. In our experience, disagreement about the proportionality of regulation often stems from miscommunication about the capability of the AI systems being discussed. A more granular way to explain the capability of a model may help better understand how to regulate it appropriately.

[20] Hendrycks et al. (2023). *An Overview of Catastrophic AI Risks*; https://arxiv.org/pdf/2306.12001.pdf
[21] National Institute of Standards and Technology. (March 22, 2023). *The Language of Trustworthy AI: An In-Depth Glossary of Terms*

Practical actions Australia can take

This chapter explores seven immediate actions that the Australian Government could take to tackle potential catastrophic and existential risks from advanced AI.

1. **Join other countries and experts by acknowledging the risk.**
2. **Ensure that any risk-based approach to regulation "scales up" to meet these kinds of extreme risks.**
3. **Show international leadership, both on global AI governance via multilateral forums and in standards development and other conformance infrastructure.**
4. **Support AI safety research and make AI safety products and services a key Australian export. Follow other nations by launching a national or regional AI technical laboratory.**
5. **Ensure a risk-based approach to regulating AI is agile and appropriately scales to high-risk advanced AIs. We provide a 5-step example to understand what this could look like.**
6. **Ensure Australians have access to justice by creating joint culpability between developers and deployers for the harms of AI. Responsibility for harm must sit with those best able to prevent it.**
7. **In light of the urgency, breadth and complexity of the issue, adopt an approach to governance that is proportionate to the challenge.**

Acknowledge the catastrophic and existential risks of AI

The first step the Australian Government must take is to acknowledge the possibility of catastrophic or existential risks from advanced AI systems. Doing so would align the Australian Government with many AI experts, AI companies, and international political leaders, and unlock a first step towards addressing the risks through effective policy and governance.

Generally "catastrophic risks" refer to those that could damage human wellbeing on a global scale or endanger civilisation.[22] Other catastrophic risks include pandemics and nuclear war. Existential risks are those that threaten the premature extinction of humanity or the permanent destruction of the potential for a desirable future.[23]

Expert assessments of catastrophic and existential risk

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" has recently been signed by the heads of companies that are developing the most advanced AI systems (OpenAI, Google DeepMind, Anthropic, Stability AI) as well as world-leading academic and other AI researchers.[24]

Two of the signatories were Geoffrey Hinton and Yoshua Benigo, both computer scientists who shared the 2018 Turing Award – similar to a Nobel Prize – for deep learning, and are popularly known as the "Godfathers of AI", along with a third recipient, Yann LeCunn.

Geoffrey Hinton quit Google Brain in May 2023 so he could "freely speak out about the risks of AI".[25] He has since described AI as an "existential threat" and described

[22] Martínez & Winter. (2022). *Ordinary meaning of existential risk.* LPP Working Paper No 7-2022. http://ssrn.com/abstract=4304670

[23] Cotton-Barratt & Ord. (2015). *Existential Risk and Existential Hope*. Future of Humanity Institute – Technical Report #2015-1. https://www.fhi.ox.ac.uk/Existential-risk-and-existential-hope.pdf

[24] Center for AI Safety. (2023). *Statement on AI risk.* https://www.safe.ai/statement-on-ai-risk

[25] Metz, C. (1 May 2023). *'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead.* https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

how powerful AI systems could be misused to cause harm, or how systems could manipulate or replace human intelligence.

Yoshua Bengio has been optimistic about the benefits of AI over his 40-year career, but has recently changed his views, writing an extensive article explaining his views about the catastrophic risks of AI, especially what he describes as “superhuman AI”, which is a system that outperforms humans on many tasks.[26] This is akin to what this paper calls “advanced AI”. In a recent ABC News article, Bengio estimated a 20% likelihood of catastrophe from superhuman AI systems being misused or acting outside human control, and expected superhuman AI systems to be developed within 5-20 years.[27]

Political leaders acknowledge catastrophic and existential risks from AI

In concert with calls from experts in AI, world political leaders have publicly acknowledged the catastrophic and existential risks from AI, and have in some cases committed to addressing these risks.

For example, the Prime Minister of the United Kingdom, Rishi Sunak, discussed “existential risks” from “superintelligent AI” in a meeting with the CEOs of Google Deepmind and OpenAI in May 2023,[28] and subsequently announced a £100 million taskforce for safe and reliable AI models.[29]

The Secretary-General of the United Nations, António Guterres, acknowledged that “the alarm bells over the latest form of AI are deafening”.[30] Similarly, at the first United Nations Security Council to discuss global governance of generative AI like chatGPT and other LLMs and multimodal foundation models

[26] Bengio, Y. (2023). *FAQ on Catastrophic AI Risks.* https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/

[27] ABC News. (2023). *AI’s dark in-joke*. https://www.abc.net.au/news/2023-07-15/whats-your-pdoom-ai-researchers-worry-catastrophe/102591340

[28] Hern, A. & Stacey, K. (25 May 2023). *No 10 acknowledges ‘existential’ risk of AI for first time*. https://www.theguardian.com/technology/2023/may/25/no-10-acknowledges-existential-risk-ai-first-time-rishi-sunak

[29] UK Government (2023). *Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI*. https://www.gov.uk/government/news/initial-100-million-for-expert-taskforce-to-help-uk-build-and-adopt-next-generation-of-safe-ai

[30] United Nations. (2023). https://press.un.org/en/2023/sgsm21832.doc.htm

(MFM), several delegates described the importance of acknowledging and addressing catastrophic and existential risks.[31]

Overall, the fact that the world is currently struggling with pressing risks from current AI systems is a clear indication that we are dramatically unprepared for the rapid acceleration and potential step-change of those risks in the near future.

Given the gravity of the risks, and their implications for the safety of the Australian public, the Government must immediately take the first step of acknowledging the possibility of catastrophic and existential risks from AI. Clear national leadership is the first step to tackling the problem.

Take a broad approach to addressing risks from AI

The Australian Government's risk-based approach to AI must go beyond the present risks of existing AI systems. The rapid pace of AI development and deployment means that a risk-based approach focused on addressing only current hazards, exposure, and vulnerability will quickly become inadequate for protecting Australians.

This is because new advances in AI will change the inherent *hazards* posed by the technology; the widespread deployment or embedding of AI into many sectors of the economy and society will increase the potential for *exposure* to those harms; and regulatory action calibrated to address only current pressing harms will lead to *vulnerability* through insufficient adaptability to the changing context of risk.

This means that pressing risks from current AI systems, such as dis/misinformation and algorithmic bias, can be intensified by further advances in AI capabilities, and widespread deployment of AI systems with these capabilities. One example would be that AI-generated image, audio, or video content ("deepfakes") become indistinguishable from authentic recordings. Another example would be supercharged scams due to message personalisation, human-level or superhuman persuasion and manipulation capabilities of the next AI models.

[31] United Nations. (2023). https://press.un.org/en/2023/sc15359.doc.htm

However, the Australian risk-based approach must also account for novel sources of risks from new AI models, and the misuse or unintended use of dangerous emergent capabilities from these models. This could involve self-replicating autonomous systems that are used for cyberwarfare or misinformation campaigns; the design of dangerous technology such as engineered pandemics or weapons; and misaligned or "rogue" AI systems that act in conflict with the intent of designers and users.

When we discuss the potential for catastrophic or existential harms from AI, we include both intensification of current pressing risks, as well as novel sources of risks.

Sources of catastrophic risks from AI

The sources of risk from AI are multidimensional, and so require a range of responses, coordinated nationally and globally.

A recent report by the Centre for AI Safety usefully summarised AI risks into four dimensions:[32]

> **Malicious use:** where an actor intentionally harnesses advanced AI to cause widespread harm. The authors specifically identify risks such as: proliferation of bioterrorism capabilities; the deliberate dissemination of uncontrolled AI agents; and the use of AI capabilities for propaganda, censorship, and surveillance.
>
> **AI race dynamics:** without appropriate intervention, competitive pressure will cause nations and corporations to rush the development of AI, and progressively cede control to AI systems. Militaries might face pressure to engage in automated warfare, where accidents can spiral out of control before humans have the chance to intervene. Corporations will face similar incentives to deploy AI to automate human labour while prioritising profits over safety. This may result in widespread dependence on unsafe AI systems.
>
> **Organisational risks:** Disasters such as NASA's Challenger Space Shuttle explosion show that, despite ample talent, time and funding, the worst

[32] Hendrycks et al. (2023). *An Overview of Catastrophic AI Risks*. https://arxiv.org/abs/2306.12001

outcomes are still possible. The vulnerability of organisations to make critical errors in high-stakes circumstances indicates that similar errors may occur when developing advanced AI systems; this has the potential to cause immense harm in unpredictable ways.

Rogue AI: A situation in which humanity loses control of an advanced AI system. Anyone who has worked with machine learning knows how it can be highly capable, yet frustratingly unpredictable. Most advancements in AI arise from trial and error, with results difficult or impossible to explain. With autonomous AI systems, these issues become much worse, manifesting as deceptive tactics that pursue rewards through shortcuts and manipulation. Most concerningly, researchers are warning that it will be difficult to prevent an advanced AI from pursuing additional power to achieve its goals. This risk is amplified where race dynamics cause corporations and militaries to give additional capabilities to AIs.

How to address catastrophic risks through regulatory and non-regulatory actions

Diverse risks will require diverse mitigations and action from a broad range of Government agencies. For example:

- Preventing **malicious use** of advanced AI will require an evolution in law enforcement capabilities.[33] There is a well-known need for greater capacity to identify and combat malicious uses, but less widely discussed is the need to coordinate with international authorities to prevent the proliferation of dangerous AI capabilities,[34] which can arise even in seemingly benign applications.[35]

- Curbing **AI race dynamics** will require strong domestic regulation to prevent unsafe "races" between Australian-based companies, combined

[33] M. Brundage et al. (2018) *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.* https://arxiv.org/abs/1802.07228

[34] Anderljung et al. (2023) *Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?* https://arxiv.org/abs/2303.09377

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

with global leadership to prevent similar dynamics globally.[36] An exemplary approach to regulating AI within our own borders will be an essential starting point.

- Countering **organisational risks** will require fine-detailed industry regulations that are pre-emptive, air-tight and adaptive. Regulators will need to coordinate the development of international safety standards that take safety seriously,[37] as well as a high-calibre assurance ecosystem that can monitor adherence with those standards at each stage of the AI lifecycle.[38, 39] Substantial penalties must be established for non-compliant organisations.[40]

- The potential for **rogue AI** presents a novel set of challenges for authorities. An international research effort into AI safety will be the first step towards solving these challenges, and this should be a key focus of Government's response. In the long-term, sophisticated law-enforcement capabilities will also need to be developed and maintained to combat this new risk, which constitutes an entirely new frontier for agencies like ASIO, ASD and the AFP.[41] If things trend badly in the coming months and years, Government should update the Australian Government Crisis Management Framework (AGCMF) to address a rogue AI crisis and regularly exercise its response, including with international partners.

The picture presented here of rapidly advancing capabilities leading to the intensification of pressing risks and novel sources of risk is focused on the potential harms from unsafe AI. We acknowledge the great potential benefits of advanced or highly capable AI systems, and recognise that Australia should be positioned to realise these benefits. However, these benefits can only be realised if appropriate arrangements are in place to safely shape these transformations,

[36] Brundage et al. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.* https://arxiv.org/abs/1802.07228
[37] Ho et al. (2023). *International Institutions for Advanced AI;* https://arxiv.org/abs/2307.04699
[38] Mokander et al. (2023). *Auditing Large Language Models: A Three Layered Approach.* https://arxiv.org/abs/2302.08500
[39] Raji et al. (2022). *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance.* https://arxiv.org/abs/2206.04737
[40] Anderljung et al. (2023) *Frontier AI Regulation: Managing Emerging Risks to Public Safety.* https://arxiv.org/abs/2307.03718
[41] Hendrycks et al. (2022). *X-Risk Analysis for AI Research.* https://arxiv.org/abs/2206.05862

with a clear-eyed and comprehensive risk-based approach that includes catastrophic and existential risks from AI.

Demonstrate international leadership

Efforts to reduce harms in Australia, while essential, will be ineffective if similar measures are not adopted internationally.[42]

A key concern is that the US and China could be engaging in a dangerous arms race. This is typified by a recent comment by international peace experts:

> "One of the questions we get most frequently from officials in Washington is: "*Who's winning the U.S.-China AI race?*" The answer is simple and unsettling: Artificial intelligence is winning, and we're nowhere near ready for what it will bring."[43]

Equally, there are emerging positive trends that Australia should vocally support.[44] At a meeting of the UN Security Council on 18 July 2023, the Chinese delegate Zhang Jun said:[45]

> The international community needs to… ensure that risks beyond human control don't occur… We need to strengthen the detection and evaluation of the entire lifecycle of AI, ensuring that mankind has the ability to press the pause button at critical moments.

This is an invitation from China to identify advanced indicators of catastrophic or existential risks and have a globally enforceable mechanism to "pause" while risks are understood and resolved. Similarly, the US is taking globally significant action to require security testing by internal and external experts before systems are released.[46]

[42] Ho et al. (2023). *International Institutions for Advanced AI.* https://arxiv.org/abs/2307.04699
[43] Cuellar & Sheehan. (2023). *AI is Winning the AI Race.* https://foreignpolicy.com/2023/06/19/us-china-ai-race-regulation-artificial-intelligence/
[44] United Nations. (2023). *International Community Must Urgently Confront New Reality of Generative, Artificial Intelligence, Speakers Stress as Security Council Debates Risks, Rewards.* https://press.un.org/en/2023/sc15359.doc.htm
[45] Recording of the 18 July 2023 meeting of UN Security Council. https://www.youtube.com/watch?v=ae7A8sQE7wg&t=56m11s
[46] BBC. (2023). *Seven AI companies agree to safeguards in the US.* *https://www.bbc.com/news/technology-66271429*

Australia is well-positioned to seize on these opportunities. For instance, Australia could:

- Take up China's offer, and support it and the US to identify advanced indicators of risk (such as advanced AIs that act autonomously, are incorrigible, and attempt to seek power using deception) and build a global "pause" mechanism (e.g. a treaty where signatories agree to freeze more advanced AI development if these advanced indicators occur).

- Demonstrate that it is a model global-citizen by agreeing to implement into domestic law the best practices adopted internationally.

The UK is an example of a country that has already shown impressive agility in its attempts to coordinate efforts internationally, meeting with leaders from AGI labs and coordinating what may be the first Governmental summit for AI safety.[47] Australia could significantly influence international governance by attending and encouraging shared commitments and consensus statements from attendees; discussing the shape of new international institutions for research and coordination (similar to the IPCC for climate change), calling for commitments from leading AI organisations, and legitimising the need to acknowledge and address risks from AI.[48]

The pervasive effect of industry lobbying means that Australia adding its voice to international efforts may make the difference. For instance, when the EU adjusted its AI Act to respond to emerging concerns from advanced AI, it later wilted under pressure from OpenAI, softening its regulations to exempt its products from requirements such as transparency, traceability, and human oversight.[49] The Australian Government has a track record of holding large technology companies to account for how their actions affect Australians;[50] AI is another opportunity for Australia to signal that the fox shouldn't set the rules for the hen house.

[47] Guardian. (2023). *Rishi Sunak's AI summit: what is its aim, and is it really necessary?* https://www.theguardian.com/technology/2023/jun/09/rishi-sunak-ai-summit-what-is-its-aim-and-is-it-really-necessary

[48] Garfinkel & Heim. (June 2023). *What Should the Global Summit on AI Safety Try to Accomplish?* https://www.governance.ai/post/what-should-the-global-summit-on-ai-safety-try-to-accomplish

[49] Time. (2023). *Exclusive: OpenAI Lobbied the E.U. to Water Down AI Regulation* https://time.com/6288245/openai-eu-lobbying-ai-act/

[50] ACCC. (2021). *New media bargaining code*. https://www.accc.gov.au/by-industry/digital-platforms-and-services/news-media-bargaining-code/news-media-bargaining-code

Immediate and practical action could focus on AI standards development. Australia should ensure that safety is a top priority for all Australians involved in standard development and negotiation. What safety looks like will vary based on the content of the standard, but ensuring AI is able to be understood by humans will typically be good, as will ensuring future systems can be subject to measurement and assessment for safety-relevant factors, such as how robust, corrigible or biased they are and whether they can be subject to dual-use.

Support AI safety research

Since it emerged as a field of research, AI safety has suffered from a severe lack of both supply and demand. Importantly, improving AI safety requires both technical and non-technical research that addresses the socio-technical system of risk and how to reduce harms through understanding and reducing hazards, exposures, and vulnerabilities in these complex systems.[51] The following recommendations are focused on how Australia can mitigate risks and reap benefits by investing in the *supply* of AI safety.

The primary reason that technology companies are currently opposing safety regulations is that they have severely under-invested in making their products safe. To date, the general absence of regulation and lack of market incentive has left AI safety research neglected and struggling to scale relative to capabilities research.

The mismatch of safety relative to capability will continue unless governments act. For example, Microsoft has impressive AI research capabilities, and yet its failed efforts to implement safety measures when rolling out Bing Chat seemingly showed either negligent disregard or an inability to implement well-understood techniques.[52]

OpenAI appears to be solidifying AI safety as a secondary concern. As part of its mission to create “superintelligence”, it has allocated just 20% of its compute

[51] Hendrycks et al. (2023). An Overview of Catastrophic AI Risks. https://arxiv.org/abs/2306.12001

[52] Marcus, G. (2023). *Why *is* Bing so reckless?* https://garymarcus.substack.com/p/why-is-bing-so-reckless

budget to AI safety. This is a questionable decision given how computationally intensive their controversial AI safety research agenda will be.[53, 54]

The tech industry needs to adapt its approach, and introducing firm regulations will be an important step in stimulating demand for AI safety. Increased demand will create a corresponding boost in supply, and it is in this emerging industry that Australia has the opportunity to establish itself as a leader.

At this moment, the AI safety industry is nascent but could grow rapidly. When OpenAI engaged the Alignment Research Centre to conduct safety audits on GPT-4, it foreshadowed the emergence of an industry.[55] The recent announcement by the Biden Administration that companies have committed to internal and external security testing of AI systems before their release has solidified this direction.[56]

This is an opportunity for Australia because:

- Strategic investment and coordination will be the determining factors of success in the emerging field of AI safety. Australia has the capability and capacity to act now.

- AI safety products and services can be easily exported, meaning Australia could find a valuable market niche and our geography is no barrier.

- The Australian Government is well placed to actively support its domestic AI safety industry by advocating for regulations internationally, funding universities, and providing tailored support and advice to its domestic providers.

[53] Snoswell, A. (2023). *What is 'AI alignment'? Silicon Valley's favourite way to think about AI safety misses the real issues.* https://theconversation.com/what-is-ai-alignment-silicon-valleys-favourite-way-to-think-about-ai-safety-misses-the-real-issues-209330

[54] OpenAI. (2023). *Introducing Superalignment.* https://openai.com/blog/introducing-superalignment

[55] Alignment Research Centre (2023) Update on ARC's recent eval efforts https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/

[56] White House. (2023). *Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI.* https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

To illustrate what this emerging industry may look like, below are specific examples of promising research directions and services that are implied by the regulations proposed in the report *Frontier AI Regulation: Managing Emerging Risks to Public Safety*:

- Adversarial testing to assess predictability and controllability
- Dangerous capabilities evaluations (e.g. via simulated environments)
- External audits to assess compliance with safe development standards
- Automated model explanation tools
- Developmental interpretability tools (i.e. explaining changes during training)
- AI activity monitoring and anomaly detection software
- Criminal AI deployment surveillance tools
- High-security software environments for exchanging and using advanced AI models

Some of these opportunities (e.g. external audits) only require safety measures to be mandated before becoming feasible and viable as an international commerce opportunity. Others, such as automated model explanation software, will require significant R&D investments.[57]

The benefits of these investments will begin with protecting our reserves of local talent. Australia has traditionally struggled to retain its elite graduates in technical disciplines, who often move overseas seeking better opportunities. However, with substantial investment into R&D, Australia can begin to retain local talent and capitalise on emerging opportunities in AI safety.

An example of this local talent is Melbourne University's Deep Learning Group. In collaboration with the Tokyo Institute of Technology, this group has quietly become a world-leader in *Developmental Interpretability*, a promising area of deep learning research that aims to quantify capabilities and risks as they arise during model development.[58]

[57] Anthropic. (2023). *Charting a Path to AI Accountability.* https://www.anthropic.com/index/charting-a-path-to-ai-accountability

[58] Murfet et al. (2023). *Towards Developmental Interpretability*. https://www.alignmentforum.org/posts/TjaeCWvLZtEDAS5Ex/towards-developmental-interpretability

Although such examples are exciting, they are also rare. As such, there needs to be a coordinated effort to fund high-quality research and train AI experts. Doing so will ensure a diverse range of products and services can be established to capture the value on offer.

Establish an Australian AI Lab as a first step

An important first step to both building Australia's AI Safety industry and supporting effective regulation is the establishment of a Government run lab focused on analysing risky systems before they operate in Australia and monitoring systems after they are deployed. In the same way the Australiasian New Car Assessment Program (ANCAP) assesses the safety features and technologies of cars, we need a regulator supported by an assessment program that can review and monitor AI systems.

Importantly, there is already a growing international precedent for this kind of "national technical laboratory" that focuses on ensuring AIs are safe and interpretable. The Tony Blair Institute has proposed a UK "AI Sentinel" to perform this function.[59] Singapore has gone a step further and created the AI Verify Foundation to focus on developing testing tools and using them to enable responsible AI.[60]

Australia could even lead a regional body, akin to the European Centre for Algorithmic Transparency.[61] ANCAP again provides a precedent for how a technical body could support not only Australia, but also our region and develop international connections with global peer organisations.

Another helpful analogy for thinking about how a regulator benefits from being supported by technical expertise and global frameworks is aviation safety. In that framework, we have a strong regulator (CASA) backed by a robust legal regime, a technical authority that reviews accidents and "near-misses" (ATSB), and a global peak body (ICAO) that interfaces with airline manufacturers, airlines, regulators

[59] Tony Blair Institute for Global Change. (2023). *A new National Purpose: AI promises a world-leading future of Britain.* https://www.institute.global/insights/politics-and-governance/new-national-purpose-ai-promises-world-leading-future-of-britain

[60] AI Verify Foundation. (2023). https://aiverifyfoundation.sg/ai-verify-foundation/

[61] European Centre for Algorithmic Transparency (ECAT). (2023). https://algorithmic-transparency.ec.europa.eu/

and technical authorities. What this regulatory regime does is give confidence to Australians that it is safe to fly without having to become technical experts themselves. This is exactly what we need for AI.

Minister Husic has indicated a potential appetite for this direction by striking an agreement with OpenAI to give Australian scientists and researchers access to OpenAI's models, including future LLMs and MFMs. Agreements like this with AI labs are the critical first step. The next step is ensuring those models are provided to a trusted evaluator and monitor who is properly connected with a functioning regulatory scheme.

Overall, viewing AI Safety as an emerging industry is a promising strategic direction for Australia. Similar to how being a first mover on historical technological advances bolstered Taiwan's and Israel's strategic influence, Australia could leverage the emerging AI safety industry to protect Australians, build an export market and secure our strategic influence in an AI-driven future.

A five-step approach to high-risk AI

The Australian Human Rights Commission, in discussing what approaches to regulation are appropriate for various kinds of technologies, draws an analogy to aviation safety (emphasis added):[62]

> Governments tend to regulate high-risk activities and technologies more closely. This helps explain the comparatively strict laws that govern fields such as gene technology, aviation, healthcare and the energy industry. In these areas, **regulation often applies both to the technology itself and how it is used**. From a human rights perspective, the need for more prescriptive regulation will be greater where the use of a specific technology carries greater risks of harm to humans.

This concept is helpful in two ways.

First, it points usefully at the existing aviation safety framework as a way to build trust in an advanced technology that is risky and user-facing. The Government could do well to apply this kind of framework – including a strong regulator, a technical body, and coordinated international governance – to AI. A further

[62] Australian Human Rights Commission. (2021). *Human Rights and Technology.* https://tech.humanrights.gov.au/sites/default/files/2021-05/AHRC_RightsTech_2021_Final_Report.pdf

strength of the aviation safety approach is that it is adaptable to changes in technology and risk rather than taking a "one and done" approach.

Second, the distinction between regulating uses and regulating technology itself is helpful in giving more nuance to a "risk-based approach". Specifically, many of today's systems might be appropriately regulated based on their possible uses, while future and advanced systems will need to be regulated as a technology regardless of their potential uses.

As set out above, advanced AI systems and their precursors are by far the greatest source of risk to humans. That means they should be the main focus of future regulatory action. In general, an **advanced AI system** is one which can:

- Complete a diverse range of reasoning tasks with human-level performance
- Navigate complex information environments and act within those environments autonomously, and
- Form sophisticated plans and reason about the consequences of actions.

These advanced systems may be able to complete a range of complex tasks with an aptitude that approaches or exceeds human capabilities. **Precursor systems** are those which can be used to help develop advanced systems, including in combination.

The five-step approach

The following is an outline of a possible five-step management process to reduce the risk of these kinds of AIs.

This proposal draws on a range of research and intends to be indicative of how long-standing risk management approaches could be applied to the risks of advanced systems and their precursors. The approach does not stand alone and would need to be connected with other proposals set out in this paper (most obviously international governance) and subject to ongoing refinement as our understanding of the technology and risks evolves.

1. **Any AI systems above a certain scale should be subject to risk assessment and classification.**

“Scale” should be measured by reference to the technical parameters of the system, rather than a measure that is distinct from the technology itself (like use case, number of users, or market size). For instance, a regulator might set this bar as “any AI system over 5 billion parameters”.

All narrow AI systems operating today would likely be below that threshold and could be subject to risk-based regulation that focuses on use cases or is industry-specific.

Systems above this threshold could be classified as “low-risk”, “precursor”, or “advanced”. “Low-risk” systems, like those below the threshold, could be subject to light touch regulation.

2. **Regulation should limit the “linking up” of precursor AI systems to ensure that advanced systems are not developed in an unmonitored way.**

The main risk of **precursor systems** comes from the possibility of them being “linked together” (as set out above) to create advanced systems. Regulations targeting precursor systems should focus on preventing that from happening.

Actions might include:

- Ensuring that providers who offer precursor systems in Australia are licensed by the regulator, are trustworthy, and understand their obligations to prevent misuse of their system.

- Requiring compute providers to ensure that significant amounts of compute (such as access to GPU clusters) are only made available to licenced providers.

- Restricting user access to systems to the minimum necessary for their business case. Most users should be confined to a user interface or API.[63] Full access to the model parameters should only be allowed where there is a demonstrated need.[64]

[63] Shevlane, T. (2022). *Structured access: an emerging paradigm for safe AI deployment.* https://arxiv.org/abs/2201.05159

[64] Anderljung et al. (2023). *Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?* https://arxiv.org/abs/2303.09377

- Requiring "upstream" providers (such as developers) to monitor for misuse by "downstream" users (such as deployers or users). This should include an obligation to prevent misuse and an obligation to report incidents to the regulator.

- Holding developers accountable to these conditions, including significant consequences if a precursor system is found to have been used to develop advanced systems without approval.

3. Advanced systems should be subject to similar regulations as precursor systems, with an additional risk assessment process conducted in collaboration with the regulator before advanced systems are developed or deployed.

Advanced systems could empower their users to cause catastrophic harm or be inherently unsafe. The main purpose of the risk assessment process is to ensure systems are not developed or deployed unless they are safe.

The risk assessment process would span both system development and deployment:

Pre-development: the developer submits a project plan, risk mitigation strategy, and summary of safety-critical technical details of the system to the regulator for approval.

Post-development: the developer and regulator collaborate to assess dangerous capabilities,[65] conduct red-teaming,[66] perform third-party audits,[67] and review cyber security practices that prevent leaks.[68] Appropriate cyber security practices post-development might include "structured access", such as operating the advanced system in a secure data centre overseen by the regulator and national technical laboratory.[69]

[65] Shevlane et al. (2023). *Model Evaluations for Extreme Risks.* https://arxiv.org/abs/2305.15324

[66] Ganguli et al. (2022). *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*. https://arxiv.org/abs/2209.07858

[67] Raji et al. (2022). *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance*. https://arxiv.org/abs/2206.04737

[68] Schuett et al. (2023). *Towards Best Practices in AGI Safety and Governance*. https://www.governance.ai/research-paper/towards-best-practices-in-agi-safety-and-governance

[69] Shevlane, T. (2022). *Structured access: an emerging paradigm for safe AI deployment* https://arxiv.org/abs/2201.05159

Pre-deployment: the developer enables the regulator to be satisfied that all risks identified during post-deployment are addressed and that the organisation has sufficient capabilities to meet ongoing regulatory obligations, such as detecting and responding to unsafe activity – either by downstream users or by the system itself.[70] Appropriate monitoring of unsafe pre-deployment activity might include the API recording metadata about usage and securely storing inputs and outputs for the benefit of technical evaluations, audits or law enforcement and national security agencies.

4. **Users of advanced systems should also meet strict requirements.**[71]

These requirements could include:

- Obtaining a permit by showing the regulator sufficient reason to need access to the advanced system and training necessary to use it safely (e.g. monitoring programs, technical expertise, and any requirements specific to the systems that emerged in the previous stages). This could include background checks where appropriate.

- Agreeing to comply with audit requirements, including storage of input and output data from all usage of the AI system.

5. **The regulator should undertake periodic audits**

Audits can help to ensure that both organisations (users) and providers (developers) are discharging their obligations, including adequately monitoring the system. Wrongdoing should attract penalties, including the potential loss of licence or permits.

Regulation to match the risk

Anchoring off current regulatory approaches to software is the wrong starting point for how future advanced AIs should be regulated. We are already seeing

[70] Yampolskiy, R. (2023). *Unmonitorability of Artificial Intelligence*. https://philarchive.org/archive/YAMUOA-3

[71] Anderljung et al. (2023). *Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?* https://arxiv.org/abs/2303.09377

"dual-use" risks where 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.[72] Future AIs that may have capabilities far exceeding humans and could act autonomously to cause widespread harms will be even higher risk.

Better anchor points for thinking about the regulation of advanced AIs is to consider it akin to Top Secret material or Biosafety level 4 laboratories. Government accepts that Top Secret material needs to be carefully controlled because the compromise of such material would be expected to cause exceptionally grave damage to the national interest, organisations or individuals.[73] Regulations exist around Biosafety level 4 laboratories because they handle easily transmitted pathogens that can cause fatal diseases, typically where there is no treatment or vaccine. This is the league of risk that advanced AI systems will operate in and that risk management should be proportionate to.

Ongoing monitoring also has an important place. AI is not like other software because it is not deterministic and doesn't always behave in the same way. "Jailbreaking" and "prompt-hacking" illustrate how new capabilities can be extracted from today's AI systems in unanticipated ways. Ongoing monitoring of advanced AIs will be essential to observing and reacting to these kinds of changes, even in systems thought to be safe when they were deployed.

In the framing of this document, we observed that many experts have called for a pause, moratorium or ban on advanced AI. This approach might be appropriate if a regulatory scheme like the above cannot be implemented and supported by sufficient technical knowledge and skill before advanced systems start to emerge. Similarly, a regulator may not be able to assure itself that advanced AI systems are safe for deployment with a current state of AI safety capability. This should be seen as the process working properly. Current technical regulators wouldn't allow an unsafe plane or unsafe laboratory to operate in Australia, and AI regulators may need to do the same.

[72] Soice et al. (June 2023). *Can large language models democratize access to dual-use biotechnology?* arXiv preprint arXiv:2306.03809
Eshoo. (October 2022). *Eshoo Urges NSA & OSTP to Address Biosecurity Risks Caused by AI.* https://eshoo.house.gov/media/press-releases/eshoo-urges-nsa-ostp-address-biosecurity-risks-caused-ai

[73] Attorney-General's Department. (January 2023). *Protective Security Policy Framework*. Chapter 8

The importance of linking this kind of response to global governance is discussed above in the context of China's recent comments at the UN Security Council.

Legal liability and access to justice

Important to managing both near-term and longer-term risks is ensuring Australians have access to justice, and the consequence of wrongdoing falls on those most responsible.

Current Australian law can be inconsistent when it comes to liability for creating and distributing potentially dangerous tools. To roughly illustrate the range of options in Australian law: the makers of encrypted messaging applications are rarely held liable for wrongs that their tools empower;[74] internet providers have special "safe harbour" provisions in the *Copyright Act 1968* to prevent them being held liable for infringements their tools empower; and overseas car makers are being held responsible for faulty airbags in cars they make that are later distributed in Australia.

Legal culpability for the harms of AI could be even less clear because the chain of providers is potentially longer than these more familiar examples. AI Labs (typically offshore) create a product that can be purchased by a business (potentially in Australia) and integrated into a product or service, and that service could cause harm to a user or to a third party, or be adopted by another business where it goes on to harm a user or third party.

As a matter of legal principle, the law should encourage the prevention of harm at the point where the harm is most easily addressed. If the law holds the wrong people responsible, we will fail to achieve justice and create dangerous incentives.

This leads to two conclusions:

[74] Taylor. (5 August 2022). *An0m: lawyers challenge encrypted messaging app used by AFP in global crime sting.* The Guardian. https://www.theguardian.com/australia-news/2022/aug/05/an0m-lawyers-challenge-encrypted-messaging-app-used-by-afp-in-global-sting

1. Consumers and businesses should have adequate legal recourse when their rights are violated. The many small businesses that will deploy AI systems may not have the means to compensate victims and strong market forces will drive risky AI deployment. This may create a dangerous dynamic where harm is likely, but immediate recourse is unavailable.

2. Because increasingly advanced AI systems are effectively a “black box” outside of the AI labs that make them, due diligence by the customers of AI labs is challenging or impossible (or creates additional risks). We are likely to see offshore AI developers attempting to use contracts to shift risk to Australian AI deployers. However, given AI labs are best placed to prevent harm in the training and deployment of their systems, liability for the consequences of systems must remain at least in part with the AI labs and developers.

These two factors together make a joint culpability scheme necessary. Specifically, AI Labs that make dangerous products must not be able to shift liability “downstream”, particularly where downstream deployers don’t have the capability or capacity to prevent AI products from doing harm or the financial wherewithal to compensate victims.

To provide an example, we have already seen a tragic example of a chatbot (Chai) persuading a user to end his own life.[75] 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.

Many Australian businesses, and perhaps even the Australian Government, are likely to roll out chatbots as part of their customer service offerings in the near future. 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. The law should be clear that, in an instance where a chatbot causes harm (like persuading or empowering a user to harm themselves

[75] 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/

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 Australia even if the developer is transparent with the deployer about that possibility.

We should also be clear-eyed about the quantum of penalties that may be necessary to dissuade dangerous behaviour. AI is almost certain to become a multi-trillion dollar global industry and 'arms race' like behaviour will encourage a "ready, fire, aim" mentality unless courts and regulators are given serious teeth. Government should also consider criminal liability in certain circumstances. For example, if a developer or deployer makes a product available in Australia knowing that it could cause direct harm (like persuading or empowering a user to harm themselves or others) or have dangerous dual-use capabilities (like being used to conduct cyber attacks, run scams, or instruct on making weapons), and that harm occurs, criminal sanction may be appropriate.

Establishing clear ground rules based on sound legal policy has the immediate benefit of ensuring that developers and deployers aren't encouraged to engage in risky behaviour. It will also have longer-term benefits. If the capability of AIs grows as we expect, the scale of the harm they can cause will grow as well, and the scale of legal consequences for wrongdoing must grow accordingly.

How an "AI Commission" could be the next step

AI could presage the biggest social transformation in human history, akin to or exceeding the industrial revolution. This transformation has to be shepherded by structures that are proportionate to its scale and impact. Using "business as usual" structures is both unfair to the public servants given the task and unlikely to achieve the outcomes Government seeks or the Australian public deserves.

Effective action relating to reducing the risk of AI, perhaps the most critical part of that transformation, is complex and complicated. Any response must navigate four key concerns:

1) Uncertainty is high and many key technical questions are yet to be resolved. But the problem is pressing and waiting for a certain path is not an option.

2) The speed of change is unprecedented, and we need to be ahead of it. Global governments can't only be responsive to progress by industry.

3) The pool of stakeholders is large and crosscuts almost all areas of Government, including industry, research, policing, national security, emergency management, international relations, taxation, education, law, employment, social services and more.

4) There's little room for error. While the analogy to aviation safety is helpful, in this case correcting after accidents might be too late.

The business-as-usual processes and structures of Government routinely navigate these issues – but never at the same time and while the stakes are so high. Perhaps the closest analogy is cyber security, where the Australian Government has experimented with various structures, including a Special Adviser to the Prime Minister on Cyber Security, an Ambassador for Cyber Affairs and a standalone Minister for Cyber Security. Some of these concepts are likely to translate to AI safety and governance – including focused diplomatic direction setting and issue-specific ministerial responsibility.

A practical next step might be creating an AI Commission, or similar body, that can act as a central point that is legible to the public and can both provide immediate direction and prioritisation to existing Government functions at the same time as designing enduring structures.

A topic-specific body has certain strengths as an alternative to leading this work from within a given department. First, any specific department or sphere of ministerial responsibility is too small a lens to fully consider and balance all the implications of AI-caused societal transformation. This was seen during DISR-led consultations where a large portion of attendees were from other departments and, despite that effort, significant Government interests were not represented. Second, a branch within a Department will inevitably be capacity-constrained when dealing with such a fast-moving, multi-faceted and high-stakes issue. This is no criticism of the capabilities of the Government teams working on the problem, but structures need to be proportionate in scale to the problem. Third, the ability to engage industry, academia, not-for-profits and the public in Australia and overseas will be critical for success. An AI Commission or similar body is more able to have a public-facing persona than a line area in a department.

Concluding comments

The potential for catastrophic or existential risks from AI needs to be recognised by Government, and Government needs to begin specific streams of effort to ensure those risks don’t harm Australians. The positive potential for AI is exciting, but hard work is required to ensure we get there.

While there might be tensions about how best to balance near-term risks and opportunities, actions of the kind set out in this paper targeting longer-term and existential risks can begin now with few short-term tradeoffs.