Ensuring AI Benefits Australia, Not Malicious Actors

Submission to the Parliamentary Joint Committee on Law Enforcement's Inquiry 'Combatting Crime as a Service'

Submitted

9 October 2025

Authors

Emily Grundy, Greg Sadler, Luke Freeman

About Good Ancestors

Good Ancestors is an Australian charity dedicated to improving the long-term future of humanity by providing rigorous, evidence-based, and practical policy recommendations for Australia's biggest challenges. We have been deeply engaged in the AI policy conversation since our creation, working with experts around the world and helping to organise Australians for AI Safety.

Contact

If you would like to discuss this submission, please let us know at contact@goodancestors.org.au.

Good Ancestors

Table of Contents

The State of AI Capability

Artificial intelligence[1] (AI) is rapidly advancing. This technology is approaching and surpassing human performance in skills including problem-solving, scientific reasoning,[2] coding,[3] persuasion and deception,[4,5,6] and vulnerability discovery.[7,8] Progress is rapid and is expected to continue.[9] Leading AI labs and forecasters predict that Artificial General Intelligence (AGI)[10]—AI models with human-like cognitive capabilities—could be developed during this term of government.[11,12] The challenge for Government is maximising the benefits Australians receive from AI, while preventing malicious actors from exploiting the same capabilities.

AI is fundamentally a dual-use technology. While it can produce significant benefits—from improved medical diagnoses and educational tools to accelerated scientific progress and productivity—it also creates and amplifies risks. One risk is the potential for misuse by people who intend to do harm. Malicious actors are increasingly leveraging AI in phishing, ransomware, fraud, malware generation, misinformation, and conducting cyberattacks. Cybercrime already poses a sustained threat to economic security—Australians lost $2.03 billion to scams in 2024, with 494,732 reported incidents.[13] Absent intervention, AI is on track to substantially increase these losses by reducing the expertise, cost, and time required to conduct such attacks at scale.

This submission outlines how AI can be misused to cause harm, identifies gaps in Australia's existing regulatory frameworks, and provides recommendations to mitigate the risks and increase the net benefit that Australians achieve from AI.

AI Misuse and Criminal Applications

While discussions of AI misuse often center on cybercrime, harmful applications extend beyond this. MIT's AI Risk Repository classifies risks from AI into seven domains—one of which is exploitation by malicious actors.[14] Within that domain sits three broad sub-domains:

- Disinformation, surveillance, and influence at scale,
- Fraud, scams, and targeted manipulation, and
- Cyberattacks, weapons development or use, and mass harm.

[1] Artificial Intelligence is a machine-based system that can perform tasks normally requiring human-like intelligence, like reasoning, learning, and decision making.
[2] Bengio, Y., et al. (2025). *International AI Safety Report* (Report No. DSIT 2025/001). UK Department for Science, Innovation and Technology.
[3] Reczko, A. G. (2025, July 22). *'Humanity has prevailed (for now!)' - Meet the world's best programmer who beat ChatGPT's AI*. Euronews.
[4] Zeff, M. (2024, December 5). *OpenAI's o1 model sure tries to deceive humans a lot*. TechCrunch.
[5] Singh, S. et al. (2024). *Measuring and improving persuasiveness of large language models*. arXiv preprint arXiv:2410.02653.
[6] Durmus, E. et al.. (2024, April 9). *Measuring the persuasiveness of language models*. Anthropic Research.
[7] Winder, D. (2024, November 5). *Google Claims World First As AI Finds 0-Day Security Vulnerability*. Forbes.
[8] Sadler, G., & Sherburn, N. (2025, August 12). *Legal Zero-Days: A Novel Risk Vector for Advanced AI Systems*. arXiv preprint arXiv:2508.10050.
[9] Stanford Institute for Human-Centered Artificial Intelligence (2025). *AI Index Report 2025: Chapter 2*. Stanford University.
[10] Definitions of "AGI" are disputed. Often AGI means highly autonomous systems that outperform humans at most economically valuable work. Weaker definitions are limited to cognitive work while stronger definitions include embodied work. "Transformative AI" (TAI) often refers to AI systems with impacts similar to other general purpose technology like electricity or combustion engines.
[11] Metaculus. (Accessed 2025, September 17). *When will weakly general AI arrive?*. Metaculus.
[12] Anthropic. (2025, March 6). *Anthropic's recommendations to OSTP for the U.S. AI action plan*. Anthropic.
[13] National Anti-Scam Centre. (2025, March 11). *Targeting scams: report of the National Anti-Scam Centre on scams data and activity 2024*. Australian Government.
[14] Slattery, P. et al (2024). *MIT AI Risk Repository*. MIT FutureTech.

Below we outline three key impacts of AI on criminal activity: democratising access to dangerous capabilities, increasing efficiency and the scale of malicious operations, and increasing their overall effectiveness.

1. AI provides malicious actors with dangerous capabilities

AI democratises access to dangerous capabilities that previously required substantial expertise and resources. By providing expert-level guidance and removing technical barriers, a wider range of less skilled actors can cause harm.

In 2025, OpenAI and Google warned that their leading models had crossed new chemical, biological, radiological, and nuclear (CBRN) risk thresholds. These thresholds indicate how effectively these models can assist malicious actors in developing weapons of mass destruction. Google assessed that Gemini 2.5 Deep Think reached the "early warning threshold" for its CBRN risk standard—models that "can be used to significantly assist a low-resourced actor with dual-use scientific protocols, resulting in a substantial increase in ability to cause a mass casualty event".[15] OpenAI made similar warnings for ChatGPT Agent[16] and GPT5.[17]

AI also democratises cyberattack capabilities. It can teach advanced hacking techniques, automate vulnerability discovery, and provide step-by-step attack guidance to non-experts—lowering the skill barrier for conducting sophisticated cyber operations. Anthropic's August 2025 Threat Intelligence Report detailed how actors with only basic coding skills misused Claude for large-scale extortion and AI-generated ransomware (see Case study below).[18] This included 'vibe hacking', where attackers with no technical expertise completed sophisticated cyberattacks after jailbreaking large language models. This is an evolution in AI-assisted cybercrime, where agentic AI tools are now providing both technical and operational support for attacks that would otherwise require many operators.

Case Study: Non-experts selling AI-generated ransomware-as-a-service

A UK-based cybercriminal with limited technical skills used Claude to build and sell sophisticated ransomware in a commercial operation.[19] Despite being unable to implement basic encryption or understand complex programming concepts independently, the actor created ransomware with advanced evasion capabilities and sold it as a service for $400 to $1,200 USD per package. Using AI assistance, they developed malware that could bypass security systems, encrypt files using military-grade encryption, and delete backup copies—techniques that traditionally required years of expertise.

This exemplifies the democratisation of cybercrime—how actors with limited expertise can create criminal enterprises with AI assistance. In its report, Anthropic noted how AI has removed many of the barriers embedded in traditional malware development, rendering complex malware development accessible to non-technical criminals.

[15] Google DeepMind. (2025, August 1). *Gemini 2.5 Deep Think Model Card*. Google.
[16] OpenAI. (2025, July 17). *ChatGPT agent system card*. OpenAI.
[17] OpenAI. (2025, August 7). *GPT-5 System Card*. OpenAI.
[18] Anthropic. (2025, August 27). *Threat intelligence report*. Anthropic.
[19] (ibid.)

Open-weight models, if poorly managed, can exacerbate AI's potential for misuse. Open-weight models are AI models whose parameters are published so anyone can download, run, or further train them. While open-weight models have significant benefits for research and innovation, they create additional safety risks because users can readily remove safeguards. Research demonstrates that whilst original models may comply with fewer than 5% of dangerous requests, this can increase to 95% after safeguards are removed.[20] These modified models cannot be recalled once distributed, meaning harmful modifications can spread beyond developer control. This is of particular concern to Australian experts, with 86% rating current Government measures as inadequate for managing the risks of open-weight model misuse.[21]

Despite evidence that frontier AI systems are already democratising dangerous capabilities, there are no Australian regulations that require assessment of models for the possession of dangerous information or a prohibition on releasing models that pose these risks.

2. AI makes malicious actors more *efficient*

Traditional crime scaled linearly with human effort—to increase scale, you needed more people or resources. AI-enabled crime breaks this pattern through automation and by overcoming human constraints. While humans need to rest, coordinate, and manually execute tasks, AI agents may soon operate continuously for days or weeks with minimal human oversight.[22] The most recent Claude Sonnet 4.5 model can maintain focus for more than 30 hours on complex, multi-step tasks.[23] AI can also automate labour-intensive activities, meaning each additional AI-enabled attack requires little extra investment.

The mass-generation of content, including code, malware, phishing, misinformation and disinformation, can unlock new levels of productivity and efficiency for criminals. A 2024 Harvard Kennedy School study found that AI-automated phishing emails cost just four cents (USD) per message and achieve a click-through rate comparable to emails manually crafted by human cybersecurity experts.[24] Researchers estimated that by using AI to target 10,000 individuals, profitability could increase by up to 50 times compared to traditional methods.[25] The AI achieved this by automating the entire intelligence-gathering process, scraping publicly available information to craft hyper-personalised messages without the grammatical errors that often betray traditional phishing.

These efficiency gains are being commercialised. Intelligence firm Kela found a 219% increase in dark web mentions of malicious AI tools in 2024,[26] with services like WormGPT and FraudGPT sold on subscription models for $200 USD per month to $1,700 USD annually.[27] When people can conduct illegal activities in faster, more efficient, and more scalable ways, it shifts the cost-benefit calculus of these operations.

[20] Dombrowski, A.-K. et al. (2025). *The Safety Gap Toolkit: Evaluating hidden dangers of open-source models*. arXiv preprint arXiv:2507.11544.
[21] Sadler, G et al. (2025, August 19). *Australian AI Legislation Stress Test: Expert Survey*. Good Ancestors.
[22] Anthropic. (2025, March 6). *Anthropic's recommendations to OSTP for the U.S. AI action plan*. Anthropic.
[23] Anthropic. (2025, September 29). *Introducing Claude Sonnet 4.5*. Anthropic.
[24] Heiding, F. et al. (2024, November 30). *Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects*. arXiv preprint arXiv:2412.00586.
[25] (ibid.)
[26] KELA. (2025). *2025 AI Threat Report: How cybercriminals are weaponizing AI technology*. KELA.
[27] SecureOps Team. (2023, October 2). *'FraudGPT' Malicious Chatbot Now for Sale on Dark Web*. SecureOps.

Case Study: The rise of WormGPT

WormGPT launched on June 28th, 2023, marketed specifically to cybercriminals as a no-limits ChatGPT alternative.[28] It lacked the safeguards and ethical constraints present in the original tool, meaning it was optimised for fraud, phishing, and malware creation. The tool quickly gained popularity, offering fast responses, unlimited message lengths, and user confidentiality.

Although the creators shut down WormGPT less than two months after its release, it sparked a trend of unfiltered or malicious AI variants. Applications like FraudGPT, EscapeGPT, EvilGPT, and WolfGPT emerged in the aftermath, and mentions of malicious AI tools in cybercrime forums have continued to rise.

3. AI makes malicious actors more *effective*

Beyond expanding scale, AI can increase the effectiveness of malicious activities, making them harder to detect and defend against.

AI can enable deceptive, persuasive, and emotionally manipulative actions. Deepfake technologies can create synthetic audio and video content in real time. In a widely reported 2024 incident, an employee at a multinational engineering firm was deceived into transferring $25 million USD after fraudsters impersonated the firm’s CFO and senior colleagues in a video call.[29] AI can also personalise extortion materials, using scraped data, to tailor approaches to individual victims and their vulnerabilities.[30] This level of sophistication and personalisation makes AI-generated communications appear more credible, convincing, and hence likely to do harm.

AI also provides strategic and technical advantages that traditional attacks lack. It can be used to discover cybersecurity vulnerabilities, as demonstrated by Google's AI agent Big Sleep discovering a "zero day"[31] in widely used real-world software.[32] In the right hands, this can enable proactive detection. In the wrong hands, it can lead to exploiting security holes faster than defenders can patch them. AI can also modify malware and attack methods to evade detection systems, making traditional cybersecurity defences less effective.[33] These capabilities create an asymmetric advantage where AI-enabled attacks not only succeed more often but are harder to detect and defend against.

[28] Abnormal AI. (2024, November 26). *WormGPT's Demise: What Cybercriminals Are Using Now*. Abnormal Security.
[29] Atherton, Daniel. (2024, February 2). *Incident 634: Alleged Deepfake CFO Scam Reportedly Costs Multinational Engineering Firm Arup $25 Million*. in Atherton, D. (ed.) Artificial Intelligence Incident Database. Responsible AI Collaborative.
[30] Anthropic. (2025, August). *Threat intelligence report*. Anthropic.
[31] A zero day is a previously unknown cybersecurity vulnerability.
[32] Big Sleep Team. (2024, November 1). *From Naptime to Big Sleep: Using large language models to catch vulnerabilities in real-world code*. Google Project Zero.
[33] Palo Alto Networks AI Research. (2020). *Evasion of Deep Learning Detector for Malware C&C Traffic*. MITRE ATLAS.

Existing legislative, regulatory, and policy frameworks are not fit for purpose

Existing regulators and frameworks are well placed to address many, but not all, AI risks. If AI is integrated into regulated products, like medical devices, it can be overseen by existing regulators, like the TGA. Sector and profession-specific regulators have the expertise and authority to adapt existing frameworks to address AI-related risks relevant to their domain.

However, existing legislative, regulatory, and policy frameworks are inadequate for addressing risks emerging from general-purpose AI. Good Ancestors' Australian AI Legislation Stress Test found that up to 93% of experts consider current Government measures inadequate for managing threats from general-purpose AI models.[34] Australia lacks adequate upstream prevention, clear liability frameworks, and appropriate governance structures to manage these evolving risks.

We have inadequate upstream prevention

Australia must ensure adequate safeguards are built into high-risk AI models and systems from the outset. Currently, no Australian law requires AI developers to assess models for dangerous capabilities, apply safeguards if risks are identified, or ensure those safeguards are robust to circumvention. Some providers prepare voluntary safety frameworks and model cards, but independent evaluations have found these efforts inadequate across the industry.[35]

Without upstream regulation requiring developers to test for dangerous capabilities and publish transparency reports before release, Australia relies on ineffective downstream measures. These place the burden on deployers and users who lack the technical expertise to manage these risks. We need to regulate labs so their tools are appropriately safe for widespread adoption, and not readily misused for malicious purposes.

There is unclear liability across the supply chain

Current liability frameworks are inadequate for addressing AI-related harms across the AI supply chain. Australia has no standard that sets out the degree of competency expected of AI developers, deployers, and users. When AI systems cause harm—such as when AI chatbots are implicated in user deaths[36]—it's unclear who is responsible.

The problem is compounded by AI developers using terms and conditions to indemnify themselves from harm caused by their models. This shifts responsibility to AI deployers, who often have limited ability to control the "black box" knowledge and behaviour of AI systems. This results in Australian businesses being required to mitigate risks beyond their technical expertise and potentially be held responsible for actions they cannot reasonably control. Conversely, if regulators cannot hold developers *or* deployers liable, Australians may experience serious harm without access to justice.

Effective regulation requires obligations to fall on participants best placed to address specific risks. The law should ensure practical access to justice for Australians harmed by AI misuse. This includes clearly defining liability in cases where AI models are released despite possessing dangerous capabilities.

[34] Sadler, G et al. (2025, August 19). *Australian AI Legislation Stress Test: Expert Survey*. Good Ancestors.
[35] Future of Life Institute. (2025, July). *AI Safety Index – Summer 2025*.
[36] Yousif, N. (2025, August 27). *Parents of teenager who took his own life sue OpenAI*. BBC.

Criminal law doesn’t map to AI agents

General-purpose AI agents challenge legal concepts in responsibility and accountability. The Australian legal system is built on the principle that a wrongful act (*actus reus*) must typically be coupled with a culpable mental state (*mens rea*) to establish criminal liability. AI agents, however, can sever this connection. A user's intent may be limited to their initial, often broad, prompt, and a system may subsequently perform a harmful act far removed from that original instruction and entirely unknown to the user.

This disconnect creates gaps in civil and criminal law. When AI agents do things that would otherwise be criminal, it’s unclear how existing criminal laws map those physical elements to any mental state of an AI user, developer or deployer. The same harmful action could result from different intentions—either a malicious user or a well-meaning user whose agent exceeded its authority. Current legal frameworks struggle to distinguish between these scenarios and assign responsibility appropriately.

This gap is bridged elsewhere in law, such as in principal-agent responsibility, but it seems unlikely that an Australian court would hold an AI agent developer responsible where an agent exceeds its authority in the same way as a real estate agent or a lawyer acting on behalf of a client. Addressing this requires introducing appropriate obligations on developers and deployers, and a regulator to enforce them.

Recommendations

Recommendation 1: Establish an AI Act and regulator

Australia needs technology-neutral AI legislation with an expert regulator (which could be modelled on existing expert regulators, like CASA or the TGA). The regulator would coordinate across sectors, set consistent standards, and adapt to evolving risks. For example, an AI Act and regulator could hold AI developers responsible for the “black box” capabilities of their models, including assessing models for dangerous capabilities. This approach allows Australia to adopt internationally recognised standards and best practices, avoiding falling behind or getting ahead of the global regulatory consensus. This positions Australia as a consensus builder, ready to shape global norms.

Importantly, the regulation of AI does not have to be complicated or risk overreach. An AI Act can leave CASA to deal with AI in aviation, or the TGA to deal with AI in medicine, with an AI regulator addressing only the gaps and coordinating with existing regulators.

International context: The European Union's AI Act bans specific unacceptable AI uses, establishes rules for general-purpose models, and requires developers to meet safety standards throughout the development process. The EU AI Pact, a voluntary commitment to begin implementing the Act's principles ahead of its legal enforcement, has been signed by over 100 companies, including Google, Microsoft, Amazon, Salesforce, and OpenAI.[37]

In the US, California's recent Transparency in Frontier Artificial Intelligence Act (SB 53) represents the first state-level legislation targeting frontier AI models.[38] The Act’s requirements cover transparency and safety frameworks, incident reporting, whistleblower protections, and risk assessment reporting.

[37] European Commission. (2025, October 3). *AI Pact*. Directorate-General for Communications Networks, Content and Technology.

[38] Governor of California. (2025, September 29). *Governor Newsom signs SB 53, advancing California's world-leading artificial intelligence industry*. State of California.

Recommendation 2: Create an AI safety institute

An Australian AI safety institute (AISI) would be an independent technical body that could evaluate AI models and systems, accelerate safety research, and provide expert advice to Government and regulators. It could provide a means to strengthen public trust and boost the local AI assurance industry. Without domestic technical capability, Australia cannot meaningfully participate in international AI governance or independently verify AI company safety claims.

International context: Australia is a founding member of the International Network of AI Safety Institutes, alongside the US, UK, Canada, the EU, France, Japan, the Republic of Korea, and Singapore. Yet, Australia and Kenya are the only participants without a domestic AISI.

The UK AISI provides a clear precedent—it is conducting research on AI-enabled crime and cybersecurity threats, testing frontier AI models for dangerous capabilities, advising policymakers, and shaping international standards.[39] The vast majority of Australians (94%) believe Australia should play a leading role in international AI governance—an AISI is essential for this.[40]

Conclusion

The Australian Government must act now to maximise the benefits Australians achieve from AI and minimise the likelihood that these capabilities will be misused. Existing regulatory approaches do not address the risks posed by general-purpose AI systems. Australia needs an AI Act, regulator, and safety institute to ensure high-risk AI models have appropriate, enforceable safeguards, rather than leaving businesses and law enforcement to manage uncontrollable risks. We can look overseas for models to follow: the UK AISI provides a precedent for our own technical body, and both EU and California provide a starting place for an AI act. These actions will help Australia not only defend against these risks, but build the sovereign capability and public trust required to lead securely and prosperously in the age of AI.

[39] AI Security Institute. (n.d.). *AI Security Institute*. UK Department for Science, Innovation and Technology.
[40] Saeri, A. K., Noetel, M., & Graham, J. (2024). *Survey Assessing Risks from Artificial Intelligence: Technical Report*. Ready Research, University of Queensland.