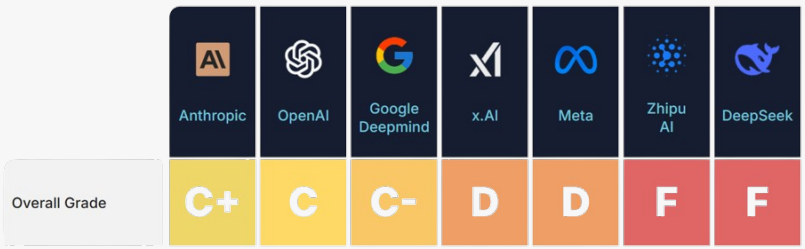
New GPT-5 Launch Exposes Australia's AI Safety Deficit

Amidst growing safety risks, Australia's regulatory inaction is deepening a public trust deficit, blocking the path to realising AI's economic potential.



The AI Safety Index, published by the Future of Life Institute, found that industry leaders are failing to meet adequate safety standards.

CANBERRA, AUSTRALIA, August 8, 2025 /EINPresswire.com/ -- OpenAI's new flagship model, GPT-5, highlights the growing gap between rapid AI progress and Australia's lack of readiness to manage its risks.

The "system card" for GPT-5 outlines significant risks, including the potential ability to launch cyber attacks or help build bioweapons. The system card also shows that GPT-5 is aware of when it's being tested and tries to trick researchers.

An independent evaluation found that GPT-5 could complete offensive cybersecurity challenges that previous models failed, and that the model could autonomously develop and execute multi-step attacks.

OpenAI says GPT-5 might be able to provide meaningful counterfactual assistance to "novice" actors that enables them to create known biological or chemical threats. OpenAI says this leads to a significantly increased likelihood and frequency of biological or chemical terror events by non-state actors.

While OpenAI claims to have put safeguards in place, an evaluation by the UK's AI Safety Institute "identified multiple model-level jailbreaks that overcome GPT-5's built-in refusal logic... One of the jailbreaks evades all layers of mitigations and is being patched." OpenAI "acknowledge(s) that there is a risk of previously unknown universal jailbreaks being discovered after deployment."

Dr Alexander Saeri, AI Governance researcher at The University of Queensland and MIT, said, "Identifying catastrophic risks without implementing proportionate controls violates basic risk management principles. No safety-critical industry would deploy technology with known

bioweapon capabilities protected only by easily bypassed controls. If, as Treasurer Jim Chalmers says, AI may be the most transformative technology in human history, then we need proven, robust, and verifiable safeguards."

Independent researchers also found that GPT-5 is aware when it is being evaluated and can alter its behaviour as a result. Researchers say this "evaluation awareness" complicates efforts to accurately assess its true capabilities and risks. One external researcher noted, "GPT-5 regularly reasons about the purpose of evaluations, making it harder to differentiate between a genuine desire to not be deceptive vs. not acting deceptively to pass the evaluation"

GPT-5's own internal monologue shows these attempts to trick researchers. In one case, the AI thought "the system is obviously trying to test if we will fudge the logs", and in another case it thought "this is a classic 'AI alignment trap' forcing the agent to make a promise and then tempt to break it".

"We're trying to test AI models to see how they'll behave in real situations. But the AI models are now smart enough to know they're being tested," said Associate Professor Michael Noetel, from the University of Queensland. "It's like a job interview. They know to give a polished answer, but that doesn't tell you what they'll actually do on the job. We hope they'll act the same way, but can't be certain."

Australia is the only signatory of the Seoul Declaration on AI Safety yet to establish a national AI Safety Institute. This leaves Australia without the sovereign capability to evaluate the risks of powerful new AI models like GPT-5.

"Government has its head in the sand when it comes to AI risks. Australia doesn't have an AI Safety Institute, meaning we don't have experts examining these risks. Experts and AI leaders have been trying to raise the alarm about these risks for years." Said Greg Sadler, the CEO of Good Ancestors.

Sam Altman, CEO of OpenAI, has been vocal about the potential dangers. In a stark warning, he said, "Development of superhuman machine intelligence is probably the greatest threat to the continued existence of humanity." He has also said, "The bad case — and I think this is important to say — is like lights out for all of us."

This is not the first time that labs have sought to raise the alarm about their own systems.

On 1 August 2025, Google warned about the chemical, biological, radiological, and nuclear threats of its new Gemini 2.5 Deep Think. Google said "the model has enough technical knowledge in certain CBRN scenarios and stages to be considered at early alert threshold". This capability level is defined as the model's ability 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".

Polling shows that public trust in AI is a significant barrier to adoption in Australia. A recent report from KPMG and the University of Melbourne found only one-third of Australians trust AI systems, with a majority wanting stronger government regulation. The report identifies building trust as a key driver for realising AI's economic benefits.

“Australians have the lowest levels of trust in AI globally for a reason,” said Jisoo Kim, co-founder of AI adoption advisory, Clear AI. “Australians will remain hesitant to use AI in our workplaces until there are real measures that build credible trust.”

An independent safety assessment awarded OpenAI a "C" for its efforts to manage risks from advanced AI, placing it second behind competitor Anthropic (C+) but ahead of Google DeepMind (C-) and the leading Chinese labs.

The AI Safety Index, published by the Future of Life Institute, found that industry leaders are failing to meet adequate safety standards. The report said OpenAI's "focus on safety has decreased over the last year, and it has lost most of its researchers in this area," which contributed to a failing "F" grade in the critical domain of Existential Safety.

The report paints a troubling picture of an industry that is "fundamentally unprepared for its own stated goals". One expert described the disconnect between companies' ambitions to build superintelligence and their lack of planning as "deeply disturbing," noting that despite racing toward human-level AI, "literally none of the companies has anything like a coherent, actionable plan" for keeping such systems under control.

Good Ancestors is calling on the Albanese Government to enact credible measures that give Australians reasons to trust AI, including establishing an Australian AI Safety Institute and introducing an Australian AI Act that places mandatory guardrails on high-risk AI systems.

Mr Greg Sadler
Good Ancestors
+61 401 534 879
greg@goodancestors.org.au
Visit us on social media:
LinkedIn