GLOBAL SHIELD.

Shaping policy. Averting catastrophe

Seven steps to AI safety in Australia

Canberra, 5 November 2025 - Today, MPs and Senators gathered to see a live demonstration of how publicly available artificial intelligence (AI) models can be used to produce real-time deepfakes, supercharge phishing emails, and even guide people through the steps necessary to make a bioweapon.

At an event hosted by Global Shield Australia and Good Ancestors in Parliament House, experts from US partner CivAI produced a live deepfake of a Senator and showed how easily voluntary guardrails on publicly available AI models can be bypassed.

Following the closed-door briefing to Members of Parliament in the morning, Global Shield hosted a similar briefing at the National Archives alongside partners CivAI and the United States Studies Centre (USSC) for the broader policy community interested in AI issues.

“Calls to wait for yet further reviews ignore the evidence that action is needed now. Australia cannot afford to repeat mistakes made with previous technological developments, such as social media, where delayed action left communities exposed and regulators playing catch-up,” said Devon Whittle, Global Shield’s Australia Director.

"The risks here go even deeper than CivAI's demonstration. AI is already helping people run scams and conduct cyber attacks. Experts warn that we could even lose control of more powerful AI models. It's great that people are excited about the AI opportunity, but we also need to do something about the risks as well” said Greg Sadler, CEO of Good Ancestors.

Global Shield Australia calls on parliament and policymakers to take these seven steps to AI safety:

- Address the unique hazards associated with AI models, including their demonstrated capabilities to deceive, self-replicate, and pursue their own goals.
- Establish monitoring and incident reporting of AI incidents in Australia, to enable the government to systemically track and respond to AI-related harms across the entire economy.
- Provide a consistent baseline for regulation of AI and resolve potential inconsistencies and uncertainties between regulatory regimes, including allocating legal responsibility for harm.
- Mandate content provenance measures and labelling at scale, a key tool to deal with harms ranging from deepfakes to disinformation.

- Establish specific security standards and obligations to prevent the exfiltration and misuse of advanced and high-risk AI models and applications by rogue actors.
- Enable assessment and certification of AI models and tools, making it easier for small businesses and users to use and trust the technology.
- Align Australia with global partners and ensure we can lead in our region, minimising compliance friction and enabling Australian businesses to access global AI assurance opportunities.

“Mandatory guardrails and monitoring and reporting obligations are critical, low-hanging fruit measures the government can enact now. These would ensure Australians can benefit from the same protections in place in markets such as the European Union and California. They would also enable Australia to join regional moves by partners such as Malaysia and Vietnam in recognising the need for action,” notes Whittle.

Hiregowdara, a US expert on AI who flew in from Berkeley, California to present the briefings in Australia, noted: “We currently rely on the goodwill of major AI companies to ensure these products are safe, and then play whack-a-mole for the problems they are creating. Concerted action by governments is needed to properly address these issues at their source.”

About Global Shield

Global Shield is an international non-profit advocacy organization dedicated to reducing global catastrophic risk. Our mission is to help governments and leaders enact and implement effective policies to reduce the risk of global catastrophe from all hazards. We’re building a future in which humanity thrives. www.globalshieldpolicy.org

Contact us

For **media inquiries** please email Marvin Meintjies: marvin.meintjies@globalshieldpolicy.org

To reach **Global Shield Australia** please email: australia@globalshieldpolicy.org