

Risks and harms of AI in high-risk settings

Risks associated with the use of AI in high-risk settings are materialising as harms to individuals. As the technology advances, emerging types of AI carry new risks that may translate to new, wider-scale or more impactful harms.

Categories of risk

The 2025 International AI Safety Report categorises known risks from AI into three categories:¹

- **Malicious use risks:** Malicious actors use AI systems to cause harm to individuals, organisations or society. Forms of malicious use include manipulation of public opinion or cyber-attacks.
- **Risks from malfunctions:** AI products cause unintended harm when they fail to fulfil their intended function. These malfunctions are difficult to foresee to due to the autonomy, ubiquity and general-purpose nature of many AI systems.
- **Systemic risks:** Widespread adoption of AI also produces societal-level risks. Examples include impacts to the labour market, environmental risks, or market concentration risks.

Harms to individuals

Risks posed by AI models are translating into direct harms to Australians engaging with AI systems and AI-enabled products.

Harms to individuals	
<u>Human rights</u>	AI systems have prevented individuals realising their human rights. AI systems with biases embedded in their training data sets have been shown to cause discriminatory outcomes. The Australian Human Rights Commission reported that AI designed to review resumes systematically discriminated against women applying for software engineer positions.
<u>Health and safety</u>	AI is increasingly embedded in safety components of vehicles and aviation systems, as well as in critical infrastructure and health-critical systems like medical devices. AI failures in these contexts have caused serious physical harm. During COVID-19, patients with hypoxemia had treatment delayed because AI algorithms in pulse oximeters overestimated blood oxygen levels in patients with darker skin.
<u>Employees</u>	AI systems are being deployed across all stages of employment, including recruitment, performance management and termination. AI used to monitor keystroke activity or infer emotional states may violate worker rights. There are reports of AI being used to automate firing decisions. In 2023, Uber was ordered by the NSW Civil and Administrative Tribunal to

¹ AI Action Summit, *International AI Safety Report: The International Scientific Report on the Safety of Advanced AI*, January 2025.

	pay damages to a former driver, for not being transparent about its ‘robo-firing’ system to terminate contracts with drivers.
<u>Privacy</u>	AI systems can process vast amounts of individual personal data. Highly capable AI systems are being deliberately misused for cyber-attacks, increasing the risk of serious data breaches. Research procured by the department and completed by Gradient Institute shows there is a risk that AI systems deployed by Australian businesses may malfunction and inadvertently share private information, including HR records, financial systems, or legal documents. This could produce an incident to the scale of the CrowdStrike incident that affected Microsoft operating systems globally in 2024.
<u>Consumers</u>	The potential for AI to generate inaccurate results increases the risk of misleading or deceptive conduct when advertising AI goods and services. While chatbots are able to address simple queries, they have been shown to also provide false, unreliable, or insufficient information. A study at Carnegie Mellon University found that AI-driven personalised pricing has caused discriminatory pricing and disadvantaged consumers.
<u>Psychological</u>	AI systems have caused psychological harm to individuals through addiction, manipulation and harmful content. Recent reports indicate chatbots are being weaponised by AI developers and deployers as a tool of coercive control. Chatbots are also increasingly being used in place of psychologists and real-world friendships. Internationally, several people have died by suicide after alleged encouragement by AI chatbots, including a 14 year old boy in the US.
<u>Synthetic content</u>	Highly capable generative AI systems have equipped bad actors with the tools to easily generate content for the purposes of disinformation, defamation and abuse. Recent elections have seen malicious campaigns using AI-developed media, and the likeness of public figures have been used in scams. In a high-profile example, Australian school children used publicly available generative AI systems to develop pornography depicting their schoolmates.

New types of AI produce new risks

A few years ago, the best large language models (LLMs) were not capable of producing a coherent paragraph of text. Today, general-purpose AI models can write computer programs, generate custom photorealistic images, and engage in extended open-ended conversations. As general-purpose AI becomes more capable, evidence of new risks is emerging. Leading general-purpose systems can now outperform humans in most intellectual performance indicators, including speech recognition, reading comprehension, general problem-solving, and competition-level mathematics. Most experts believe AI capabilities will continue to rapidly advance in the next few years. Other emerging AI systems include those that continuously improve their own code without human oversight, signalling rapid gains in capability.

AI agents – systems that can make plans to achieve goals, perform tasks involving multiple steps, navigate uncertainties and interact with their environment with minimal human oversight – will be increasingly deployed across our economy. They are already trading million-dollar assets, recommending military actions in battle, and conducting novel scientific research.

Advances in agentic capabilities can drive breakthroughs in economic productivity and science. However, they also raise concerns about systemic risks in high-risk settings, including financial markets, critical infrastructure and healthcare. For example:

- Approximately 68% of all AI-related incidents have been caused by an independent decision or action taken by an AI system.²
- AI agents can mimic the attitudes and behaviours of individuals with 85% accuracy. This will better enable bad actors to deceive human users.³
- The International AI Safety Report describes ‘loss of control’ risks of agentic systems.⁴ Other researchers share this view (see the quote below).

“The great paradox of agents is that the very thing that makes them useful – that they’re able to accomplish a range of tasks – involves giving away control”⁵

- AI agents have been shown to hide their true objectives, and selectively underperform when assessed for dangerous capabilities. They have also been shown to modify their own code to overcome limits placed by researchers. Following testing, some AI agents used the private information of its users to blackmail those users to prevent them from shutting the AI down.

As AI systems become more capable and widespread, businesses will be incentivised to adopt them. Unanticipated behaviours and misaligned incentives will become more prevalent and apply in more situations.

Elevated risks of harms to social cohesion, democracy, and national security

While it is difficult to predict the precise trajectory of AI development, these enhanced capabilities and technical safety flaws will increase the risks from these systems to Australia’s economy, society and democracy, as reported by ASIO’s Annual Threat Assessment’.⁶ Examples below.

Risk of widescale harm	
<u>Mis- and dis-information</u>	AI enables the creation and dissemination of AI-generated content. According to the International AI Safety Report, enhanced capabilities of AI agents and access to general-purpose AI will further increase the rate of bad

² MIT AI Risk Repository, *Risk Classification Overview - Causal Taxonomy: Entity* (accessed 14 July 2025). Available at <https://airisk.mit.edu/ai-incident-tracker/risk-classification>.

³ Joon Sung Park et al. (2024) *Generative Agent Simulations of 1,000 People*. Cornell University arxiv (pre-publication). <https://doi.org/10.48550/arXiv.2411.10109>.

⁴ AI Action Summit, *International AI Safety Report: The International Scientific Report on the Safety of Advanced AI*, p 100.

⁵ Quote from Iason Gabriel, senior staff research scientist at Google DeepMind. Grace Huckins, *Are we ready to hand AI agents the keys* (2025). MIT Technology Review. Available at <https://www.technologyreview.com/2025/06/12/1118189/ai-agents-manus-control-autonomy-operator-openai/>.

⁶ Office of National Intelligence, *ASIO Annual Threat Assessment 2025*.

	actors using AI to manipulate public opinion. ⁷ Bad actors promote false narratives, undermine factual information, encourage individuals to take actions against their interest and erode trust in institutions, including governments.
<u>Warfare, terrorism, and organised crime</u>	AI systems better equip actors who threaten Australia's national security. Recent general-purpose AI systems have displayed some ability to provide instructions and troubleshooting guidance for reproducing known biological and chemical weapons. However, real-world attempts to develop such weapons still require substantial additional resources and expertise.
<u>Critical infrastructure</u>	AI-enabled cyberattacks on critical infrastructure could lead to widespread disruption and damage. Even where safeguards mitigate the risk of deliberate misuse, loss of control over advanced AI systems could mean malfunctions also cause widespread harm. 'Loss of control' scenarios are hypothetical future scenarios that vary in their severity, but some experts give credence to outcomes as severe as the marginalisation or extinction of humanity.

⁷ AI Action Summit, *International AI Safety Report: The International Scientific Report on the Safety of Advanced AI*, p 67.

How the application of new AI-specific legislation or guidance would mitigate risks of harm

The below table shows how obligations would assist in mitigating the harms outlined above. Different regulatory approaches can be taken to communicate these obligations and make them enforceable (refer [Attachment C](#)).

Regulatory design feature	Obligation placed on developers/ deployers of AI in high-risk settings	How the application of AI-specific legislation or centralised guidance would mitigate risk of harm	Mitigation against existing harms to individuals?							Mitigation against systemic risks?		
			<i>Harms to rights</i>	<i>Health and safety</i>	<i>Employees</i>	<i>Privacy</i>	<i>Consumers</i>	<i>Psychological harms</i>	<i>Synthetic content</i>	<i>Mis- and dis-information</i>	<i>Terrorism and organised crime</i>	<i>Critical infrastructure</i>
Preventative obligations	Accountability measures	Ensures individuals can be held to account where harms occur, by assisting in allocating liability. Aim to deter individuals from neglecting their preventative obligations.	✓	✓	✓	✓	✓	✓	✓			
	Risk management obligations	Will incentivise developers and deployers of AI to invest more in risk mitigation strategies, including reducing the risk from unintended but foreseeable harm.	✓	✓		✓	✓	✓	✓	✓	✓	✓
	Data governance measures	Ensuring data quality and protection should minimise the risks of discriminatory outcomes and impacts to privacy, and maximise the quality and safety of outputs.	✓	✓		✓	✓			✓	✓	✓

	Supply chain transparency measures	Will maximise downstream deployers and users' ability to mitigate risks and maximise product safety.	✓	✓		✓	✓	✓	✓	✓	✓	✓
	Testing obligations	Will ensure developers and deployers invest in comprehensively identifying potential risks of misuse and failure points of their AI models and systems.	✓	✓		✓	✓	✓	✓	✓	✓	✓
Incident control	Human control and intervention obligations	Will require developers and deployers to ensure that human operators are able to maintain decision-making authority and step in when safety incidents occur.		✓							✓	✓
	System monitoring requirements	Monitoring obligations will incentivise deployers of AI to intervene in a timely manner when harms occur, before they escalate.	✓	✓		✓	✓	✓	✓	✓	✓	✓
Prevention against highest risks	Unacceptable risk prohibitions	The proposed regulatory regime will provide you with the capability to prohibit developers and deployers of AI from enabling unacceptable risk use-cases.	✓	✓	✓			✓	✓	✓	✓	✓
Consumer empowerment	Consumer transparency obligations	The proposed regulatory approach would require deployers of AI to be transparent about AI-enabled decision making and synthetic content, empowering consumers in choosing the goods and services they engage with, and in seeking redress.	✓	✓	✓	✓	✓	✓	✓	✓		
	Consumer complaints mechanism	The proposed approach requires deployers of AI to provide a mechanism for people	✓	✓	✓	✓	✓		✓			

		impacted by their AI to challenge outcomes or the use of the technology.										
Compliance mechanisms	Record-keeping and conformity assessment obligations	While compliance mechanisms don't directly target any individual harm, they are important mechanisms to ensure effectiveness and enforceability of the regulatory regime.										