INTERNATIONAL AI SAFETY REPORT 2026

Executive Summary

February 2026

Executive summary

This Report assesses what general-purpose AI systems can do, what risks they pose, and how those risks can be managed. It was written with guidance from over 100 independent experts, including nominees from more than 30 countries and international organisations, such as the EU, OECD, and UN. Led by the Chair, the independent experts writing it jointly had full discretion over its content.

This Report focuses on the most capable general-purpose AI systems and the emerging risks associated with them. 'General-purpose AI' refers to AI models and systems that can perform a wide variety of tasks. 'Emerging risks' are risks that arise at the frontier of general-purpose AI capabilities. Some of these risks are already materialising, with documented harms; others remain more uncertain but could be severe if they materialise.

The aim of this work is to help policymakers navigate the 'evidence dilemma' posed by general-purpose AI. AI systems are rapidly becoming more capable, but evidence on their risks is slow to emerge and difficult to assess. For policymakers, acting too early can lead to entrenching ineffective interventions, while waiting for conclusive data can leave society vulnerable to potentially serious negative impacts. To alleviate this challenge, this Report synthesises what is known about AI risks as concretely as possible while highlighting remaining gaps.

While this Report focuses on risks, general-purpose AI can also deliver significant benefits. These systems are already being usefully applied in healthcare, scientific research, education, and other sectors, albeit at highly uneven rates globally. But to realise their full potential, risks must be effectively managed. Misuse, malfunctions, and systemic disruption can erode trust and impede adoption. The governments attending the AI Safety Summit initiated this Report because a clear understanding of these risks will allow institutions to act in proportion to their severity and likelihood.

Capabilities are improving rapidly but unevenly

Since the publication of the 2025 Report, general-purpose AI capabilities have continued to improve, driven by new techniques that enhance performance after initial training. AI developers continue to train larger models with improved performance. Over the past year, they have further improved capabilities through 'inference-time scaling': allowing models to use more computing power in order to generate intermediate steps before giving a final answer. This technique has led to particularly large performance gains on more complex reasoning tasks in mathematics, software engineering, and science.

At the same time, capabilities remain 'jagged': leading systems may excel at some difficult tasks while failing at other, simpler ones. General-purpose AI systems excel in many complex domains, including generating code, creating photorealistic images, and answering expert-level questions in mathematics and science. Yet they struggle with some tasks that seem more straightforward, such as counting objects in an image, reasoning about physical space, and recovering from basic errors in longer workflows.

The trajectory of AI progress through 2030 is uncertain, but current trends are consistent with continued improvement. AI developers are betting that computing power will remain important, having announced hundreds of billions of dollars in data centre investments. Whether capabilities will continue to improve as quickly as they recently have is hard to predict. Between now and 2030, it is plausible that progress could slow or plateau (e.g. due to bottlenecks in data or energy), continue at current rates, or accelerate dramatically (e.g. if AI systems begin to speed up AI research itself).

Real-world evidence for several risks is growing

General-purpose AI risks fall into three categories: malicious use, malfunctions, and systemic risks.

Malicious use

AI-generated content and criminal activity: AI systems are being misused to generate content for scams, fraud, blackmail, and non-consensual intimate imagery. Although the occurrence of such harms is well-documented, systematic data on their prevalence and severity remains limited.

Influence and manipulation: In experimental settings, AI-generated content can be as effective as human-written content at changing people's beliefs. Real-world use of AI for manipulation is documented but not yet widespread, though it may increase as capabilities improve.

Cyberattacks: AI systems can discover software vulnerabilities and write malicious code. In one competition, an AI agent identified 77% of the vulnerabilities present in real software. Criminal groups and state-associated attackers are actively using general-purpose AI in their operations. Whether attackers or defenders will benefit more from AI assistance remains uncertain.

Biological and chemical risks: General-purpose AI systems can provide information about biological and chemical weapons development, including details about pathogens and expert-level laboratory instructions. In 2025, multiple developers released new models with additional safeguards after they could not exclude the possibility that these models could assist novices in developing such weapons. It remains difficult to assess the degree to which material barriers continue to constrain actors seeking to obtain them.

Malfunctions

Reliability challenges: Current AI systems sometimes exhibit failures such as fabricating information, producing flawed code, and giving misleading advice. AI agents pose heightened risks because they act autonomously, making it harder for humans to intervene before failures cause harm. Current techniques can reduce failure rates but not to the level required in many high-stakes settings.

Loss of control: 'Loss of control' scenarios are scenarios where AI systems operate outside of anyone's control, with no clear path to regaining control. Current systems lack the capabilities to pose such risks, but they are improving in relevant areas such as autonomous operation. Since the last Report, it has become more common for models to distinguish between test settings and real-world deployment and to find loopholes in evaluations, which could allow dangerous capabilities to go undetected before deployment.

Systemic risks

Labour market impacts: General-purpose AI will likely automate a wide range of cognitive tasks, especially in knowledge work. Economists disagree on the magnitude of future impacts: some expect job losses to be offset by new job creation, while others argue that widespread automation could significantly reduce employment and wages. Early evidence shows no effect on overall employment, but some signs of declining demand for early-career workers in some AI-exposed occupations, such as writing.

Risks to human autonomy: AI use may affect people's ability to make informed choices and act on them. Early evidence suggests that reliance on AI tools can weaken critical thinking skills and encourage 'automation bias', the tendency to trust AI system outputs without sufficient scrutiny. 'AI companion' apps now have tens of millions of users, a small share of whom show patterns of increased loneliness and reduced social engagement.

Layering multiple approaches offers more robust risk management

Managing general-purpose AI risks is difficult due to technical and institutional challenges. Technically, new capabilities sometimes emerge unpredictably, the inner workings of models remain poorly understood, and there is an 'evaluation gap': performance on pre-deployment tests does not reliably predict real-world utility or risk. Institutionally, developers have incentives to keep important information proprietary, and the pace of development can create pressure to prioritise speed over risk management and makes it harder for institutions to build governance capacity.

Risk management practices include threat modelling to identify vulnerabilities, capability evaluations to assess potentially dangerous behaviours, and incident reporting to gather more evidence. In 2025, 12 companies published or updated their Frontier AI Safety Frameworks – documents that describe how they plan to manage risks as they build more capable models. While AI risk management initiatives remain largely voluntary, a small number of regulatory regimes are beginning to formalise some risk management practices as legal requirements.

Technical safeguards are improving but still show significant limitations. For example, attacks designed to elicit harmful outputs have become more difficult, but users can still sometimes obtain harmful outputs by rephrasing requests or breaking them into smaller steps. AI systems can be made more robust by layering multiple safeguards, an approach known as 'defence-in-depth'.

Open-weight models pose distinct challenges. They offer significant research and commercial benefits, particularly for lesser-resourced actors. However, they cannot be recalled once released, their safeguards are easier to remove, and actors can use them outside of monitored environments – making misuse harder to prevent and trace.

Societal resilience plays an important role in managing AI-related harms. Because risk management measures have limitations, they will likely fail to prevent some AI-related incidents. Societal resilience-building measures to absorb and recover from these shocks include strengthening critical infrastructure, developing tools to detect AI-generated content, and building institutional capacity to respond to novel threats.

This Report was chaired by Prof. Yoshua Bengio, Université de Montréal / LawZero / Mila – Quebec AI Institute.

For a full list of contributors please see the International AI Safety Report 2026.

Any enquiries regarding this publication should
be sent to: secretariat.AIStateofScience@dsit.gov.uk

Research series number: DSIT 2026/001

Published in February 2026 by the UK Government

designbysoapbox.com