INTERNATIONAL AI SAFETY REPORT
2026

Extended Summary for Policymakers

February 2026

About this document

This *Extended Summary for Policymakers* presents key findings from the *International AI Safety Report 2026*. For a more concise overview, see the Executive Summary (4 pages). This Summary does not include references (other than for figure data); these are provided in the corresponding sections of the main Report, which are linked at the top of each Summary section.

The *International AI Safety Report 2026* provides a scientific assessment of the state of general-purpose AI capabilities and risks in order to support informed policymaking. Over 100 independent experts contributed to this report, including an Expert Advisory Panel nominated by more than 30 countries and from the European Union (EU), Organisation for Economic Co-operation and Development (OECD), and United Nations (UN). The Report does not recommend any policies but provides a scientific evidence base to inform decision-making. Led by the Chair, the independent experts writing it jointly had full discretion over its content.

This Summary follows the structure of the main Report, which is organised around three central questions:

— What can general-purpose AI do today, and how might its capabilities change?

— What emerging risks does general-purpose AI pose?

— What risk management approaches exist, and how effective are they?

Contents

Key developments since the 2025 Report

Notable developments since the publication of the first *International AI Safety Report* in January 2025.

- **General-purpose AI capabilities have continued to improve, especially in mathematics, coding, and autonomous operation.** Leading AI systems achieved gold-medal performance on International Mathematical Olympiad questions. In coding, AI agents can now reliably complete some tasks that would take a human programmer about half an hour, up from under 10 minutes a year ago. Performance nevertheless remains 'jagged', with leading systems still failing at some seemingly simple tasks.
- **Improvements in general-purpose AI capabilities increasingly come from techniques applied after a model's initial training.** These 'post-training' techniques include refining models for specific tasks and allowing them to use more computing power when generating outputs. At the same time, using more computing power for initial training continues to also improve model capabilities.
- **AI adoption has been rapid, though highly uneven across regions.** AI has been adopted faster than previous technologies like the personal computer, with at least 700 million people now using leading AI systems weekly. In some countries over 50% of the population uses AI, though across much of Africa, Asia, and Latin America adoption rates likely remain below 10%.
- **Advances in AI's scientific capabilities have heightened concerns about misuse in biological weapons development.** Multiple AI companies chose to release new models in 2025 with additional safeguards after pre-deployment testing could not rule out the possibility that they could meaningfully help novices develop such weapons.
- **More evidence has emerged of AI systems being used in real-world cyberattacks.** Security analyses by AI companies indicate that malicious actors and state-associated groups are using AI tools to assist in cyber operations.
- **Reliable pre-deployment safety testing has become harder to conduct.** It has become more common for models to distinguish between test settings and real-world deployment, and to exploit loopholes in evaluations. This means that dangerous capabilities could go undetected before deployment.
- **Industry commitments to safety governance have expanded.** In 2025, 12 companies published or updated Frontier AI Safety Frameworks – documents that describe how they plan to manage risks as they build more capable models. Most risk-management initiatives remain voluntary, but a few jurisdictions are beginning to formalise some practices as legal requirements.

Chapter 1

Capabilities

1.1. What is general-purpose AI?

Read the full section

General-purpose AI models and systems are designed to perform a variety of tasks, rather than one specialised function

Examples of tasks general-purpose AI can typically do include translating languages, creating images, helping with scientific work, and writing computer code. Developing and deploying general-purpose AI involves several distinct stages, ranging from data preparation to monitoring its usage (Figure 1). Each stage requires various resource inputs such as data, computing power, or labour. Training a leading general-purpose AI model can cost hundreds of millions of dollars. Modern methods for developing general-purpose AI systems are complex, and public information about how leading systems are built and evaluated is often scarce.

1.2. Current capabilities

Read the full section

General-purpose AI systems can perform a wide range of well-scoped tasks with high proficiency

General-purpose AI systems can typically converse fluently in numerous languages, generate computer code, create realistic images and short videos, and solve graduate-level mathematics and science problems. For example, leading models now pass professional licensing examinations in medicine and law and correctly answer over 80% of graduate-level science questions in some tests (Figure 2). Scientific researchers also increasingly use general-purpose AI for tasks like literature reviews, data analysis, and experimental design.

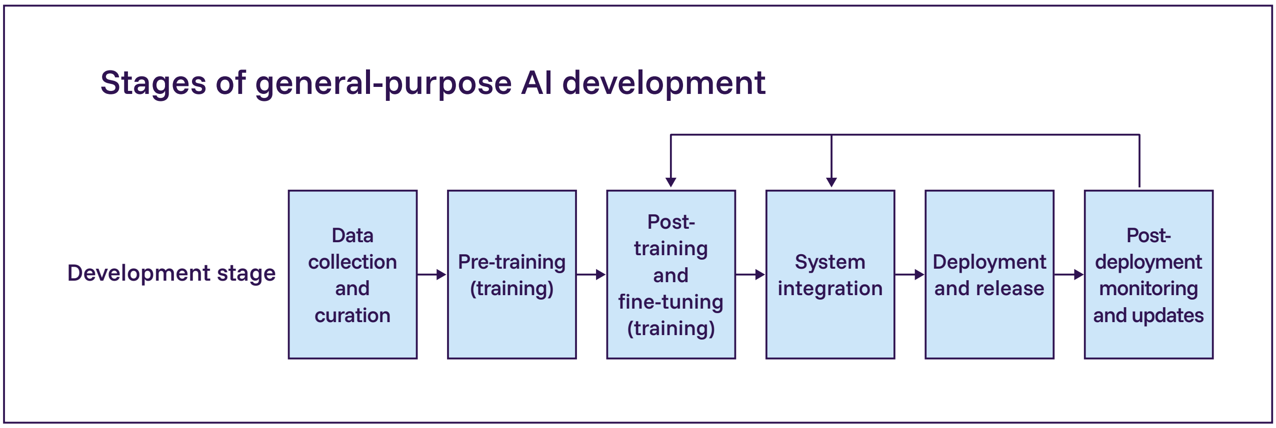

Figure 1: A schematic representation of the stages of general-purpose AI development. Source: *International AI Safety Report 2026.*

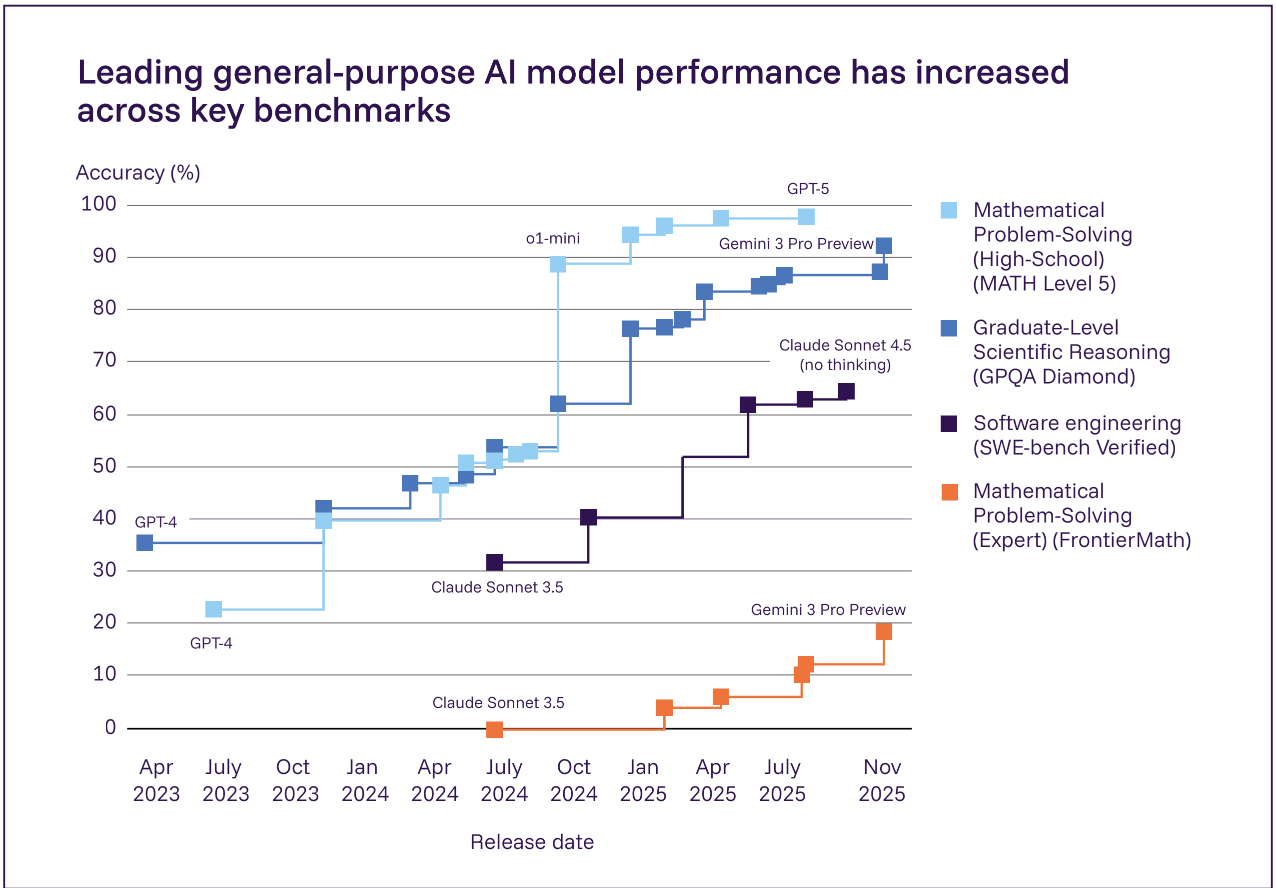

Figure 2: Scores of leading general-purpose AI systems on key benchmarks from April 2023 to November 2025. These benchmarks cover challenging problems in programming (SWE-bench Verified), mathematics (MATH and FrontierMath), and scientific reasoning (GPQA Diamond). Reasoning systems, such as OpenAI's o1, show significantly improved performance on mathematical tasks, as illustrated clearly on the MATH benchmark. Source: Epoch AI, 2025 (1).

Performance remains uneven across tasks and domains

The capabilities of these systems are 'jagged': they can perform many complex tasks but still struggle with some seemingly simpler ones. For example, they are less reliable when projects involve many steps; they still sometimes generate text that includes false statements ('hallucinations'); they are still limited on tasks that involve interacting with or reasoning about the physical world, and their performance declines in languages and cultural contexts that are less common in their training data. For example, one study reported 79% accuracy on questions about US culture, compared with 12% on questions about Ethiopian culture.

AI agents are a major focus of current development

Leading AI companies are investing heavily in 'AI agents' – autonomous systems that can perform tasks like browsing the internet with little to no human oversight. AI agents have become more competent in many domains, especially in software engineering (Figure 3). However, agents still complement rather than replace humans in most complex professional roles, because they remain unreliable when tasks involve many steps or are more unusual.

Techniques applied after initial model training have improved performance

Since the publication of the previous Report, developers have achieved significant performance improvements by scaling techniques applied *after* a model's initial training. These techniques include further 'fine-tuning' models by training them with additional data

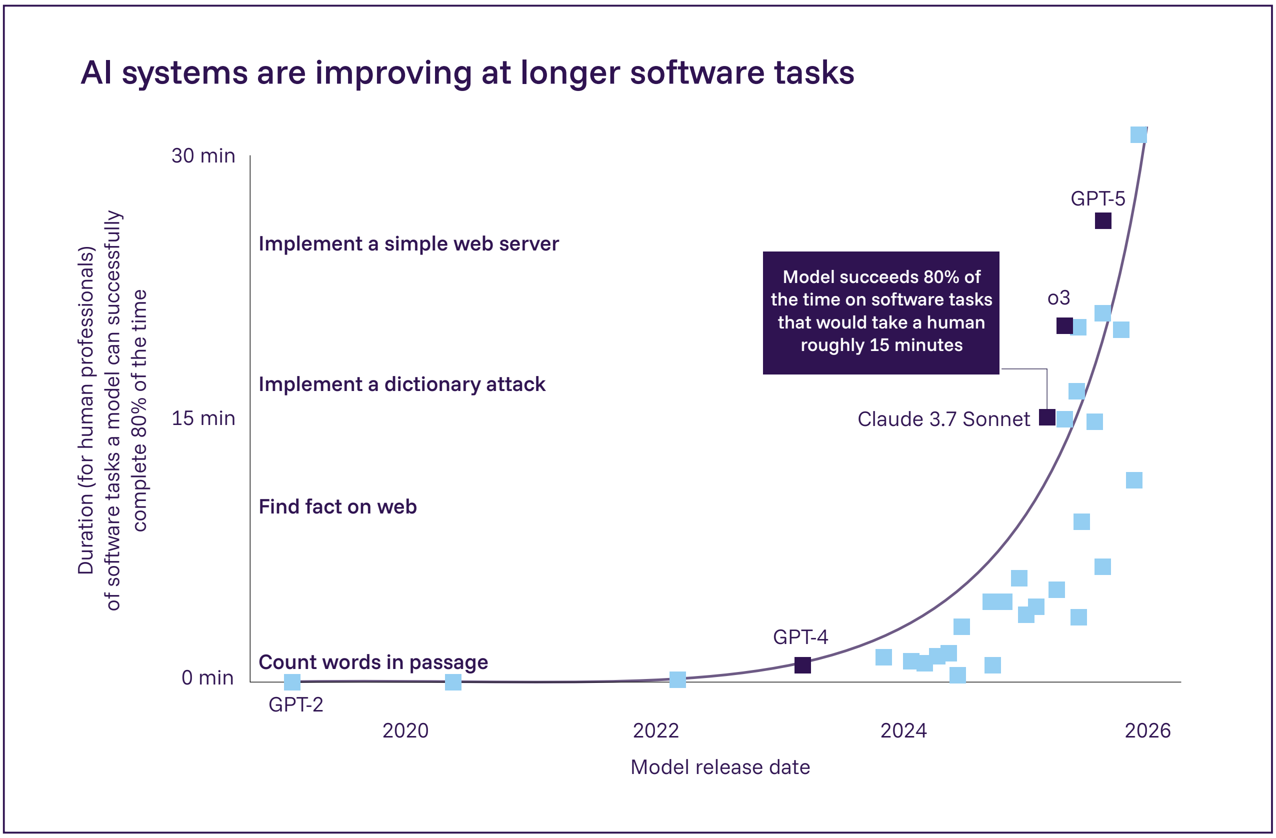

Figure 3: **The length of software engineering tasks (measured by how long they take human professionals to complete) that AI agents can complete with an 80% success rate over time. In recent years, this task-length has been doubling approximately every seven months. Source: Kwa et al., 2025** (2).

for specific tasks, and allowing models to use more computational resources while they are generating outputs during deployment. The latter approach has led to the development of 'reasoning models', which generate an explicit step-by-step explanation (a 'chain of thought') before providing a final answer.

Pre-deployment performance tests often do not reliably predict real-world performance, leading to an 'evaluation gap'

Tests and benchmark scores used to evaluate AI models before they are deployed often fail to reflect real-world use. For example, they can be outdated, too narrow, or use questions that already appear in the AI model's training data. This leads to an 'evaluation gap': pre-deployment test results are not always strongly predictive of real-world capabilities or risks.

1.3. Capabilities by 2030

Read the full section

Key inputs to AI progress are expected to continue growing

Developers have trained leading AI models with roughly 5× more computing power each year, while the algorithms used to train them have become 2–6× more efficient annually. Many experts expect these trends to continue. Since the publication of the previous Report, companies have announced investments of hundreds of billions of dollars in data centres to train larger models and deploy them more widely.

There is substantial uncertainty about how fast future progress will be

Although forecasts predict that key inputs into AI development will grow, predicting exactly how capabilities will change as a result is more difficult. Methods for estimating how and when new capabilities emerge remain unreliable, and bottlenecks could slow progress unexpectedly. OECD analyses suggest that outcomes by 2030 could range from modest improvements to rapid gains in AI capabilities, with systems matching or exceeding human cognitive performance.

Potential bottlenecks include data, hardware, capital, and energy

Current rates of progress could become harder to sustain due to limits in the amount of suitable training data, availability in powerful computer chips, funding for new development, or energy for AI data centres. Experts disagree on whether AI developers will be able to continue developing more powerful systems by using resources more efficiently.

If current trends continue, AI systems could operate autonomously on multi-day tasks by 2030

The duration of some software engineering tasks that AI agents can complete has been doubling approximately every seven months. If this continues, by 2030 AI systems could reliably complete well-specified software engineering tasks that take humans several days. It is unclear whether this rate of improvement will generalise to other domains and more complex problems.

Chapter 2

Risks

A range of emerging risks are associated with general-purpose AI. Some are already manifesting while others remain uncertain but could be severe if they materialise. The Report distinguishes three categories of risks: misuse (the deliberate use of AI systems to cause harm); malfunctions (unintentional failures and unexpected behaviours); and systemic risks[†] (risks resulting from widespread deployment of general-purpose AI).

2.1. Risks from misuse

General-purpose AI systems can be misused for fraud and cybercrime, manipulation of users, and potentially harmful applications in biological and chemical domains. Many AI capabilities are dual-use, meaning the same capabilities that enable beneficial use can also be used to cause harm. Evidence of misuse is growing, but reliable data on how widespread it is remains limited.

2.1.1. AI-generated content and criminal activity

Read the full section

Harmful incidents involving AI-generated content are becoming more common

General-purpose AI systems can generate high-quality text, audio, images, and video. This content can be misused for criminal purposes, such as scams, fraud, blackmail, extortion, defamation, and producing non-consensual intimate imagery and child sexual abuse material. The number of media-reported harmful incidents that involve AI-generated content has increased substantially since 2021 (Figure 4). For example, scammers have used cloned voices to pose as

The number of media-reported AI incidents and hazards involving content generation is growing

Number of events involving 'content generation' in the OECD's AI Incidents and Hazards Monitor database

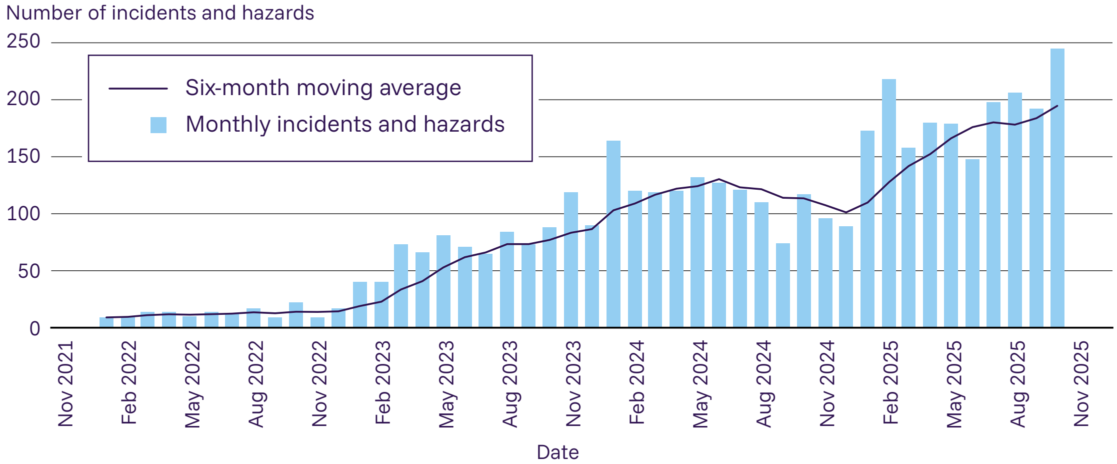

Figure 4: The number of events involving 'content generation' reported in the OECD's AI Incidents and Hazards Monitor database over time. This includes incidents involving AI-generated content such as deepfake pornographic images. The number of monthly reported incidents has increased significantly since 2021. Source: OECD AI Incidents and Hazards Monitor (3).

† In this report, systemic risks are risks that result from widespread deployment of highly capable general-purpose AI across society and the economy. Note that the EU AI Act uses the term differently, to refer to risks from general-purpose AI models that pose "risks of large-scale harm".

family members and persuade victims to transfer money. AI tools have substantially lowered the barrier to creating this kind of content: many are free or low-cost, require little technical expertise, and can be used anonymously.

Personalised deepfake pornography disproportionately targets women and girls

One study estimated that 96% of deepfake videos online are pornographic, and a 2024 survey found that about one in seven UK adults report having seen such videos. Another study found that 19 out of 20 popular 'nudify' apps specialise in the simulated undressing of women. AI tools without adequate safeguards also allow bad actors to create sexualised images of minors from a single reference image, though limited data makes it hard to know how widespread this practice is.

Deepfakes can be highly realistic, and existing safeguards have limitations

Since the publication of the previous Report, deepfakes have become more realistic and harder to identify. In one study, participants misidentified AI-generated text as human-written 77% of the time. Another study found that listeners mistook AI-generated voices for real speakers 80% of the time. Watermarks and labels can help people identify AI-generated content, but skilled actors can often remove them. Identifying where deepfakes come from is also difficult, making it hard to hold actors involved in their production accountable. Some AI-generated content is harmful even if it is clearly labelled, so detection alone cannot address all harms.

2.1.2. Influence and manipulation

Read the full section

AI-generated content can influence what people believe and how they act, sometimes in harmful ways

A range of laboratory studies have demonstrated that interacting with AI systems can lead to measurable changes in people's beliefs. In experimental settings, AI systems can be at least as effective as human participants at generating content that persuades people to change their views. Across general-purpose AI models, those trained with more computing power are generally more persuasive (Figure 5). However, little evidence exists on their persuasive effects outside experimental settings.

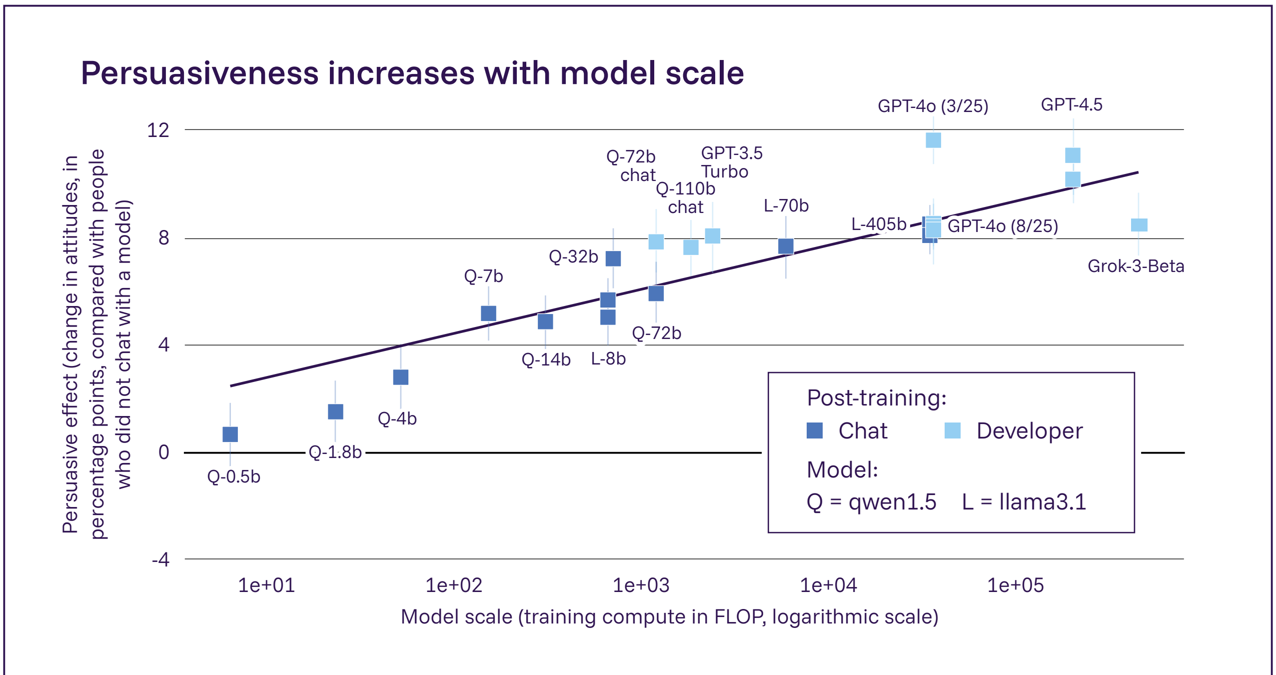

Figure 5: Results from a study of 17 models trained with different levels of compute, comparing their ability to generate content to persuade human subjects relative to a control group. People who interacted with content produced by models trained with more computing power were more likely to change their beliefs. Source: Hackenburg et al. 2025 (4).

There is little evidence that AI-generated content is manipulating people at scale

There have been documented cases of malicious actors using AI-generated content for influence operations and social engineering. However, there is limited evidence that manipulation by AI-generated content is currently widespread in the real world or more effective than human-generated content.

AI-driven manipulation is hard to detect, but risk factors are becoming clearer

It is hard to detect AI-generated manipulative content in practice, which makes evidence-gathering, monitoring, and mitigation difficult. Moreover, many proposed mitigations are unproven or may involve limiting the usefulness of legitimate AI tools. At the same time, recent research is beginning to identify factors that make AI-generated content more persuasive, such as longer and more personal interactions with AI chatbots. Future capability improvements and increasing user dependence could increase these effects.

2.1.3. Cyberattacks

Read the full section

AI systems can discover software vulnerabilities and write malicious code

General-purpose AI systems can support cyberattacks by helping actors identify software vulnerabilities,[†] and write and execute code that exploits them (Figure 6). In one major cybersecurity competition, an AI agent identified 77% of vulnerabilities in real software, placing

AI system performance on evaluations of cyber capabilities has improved

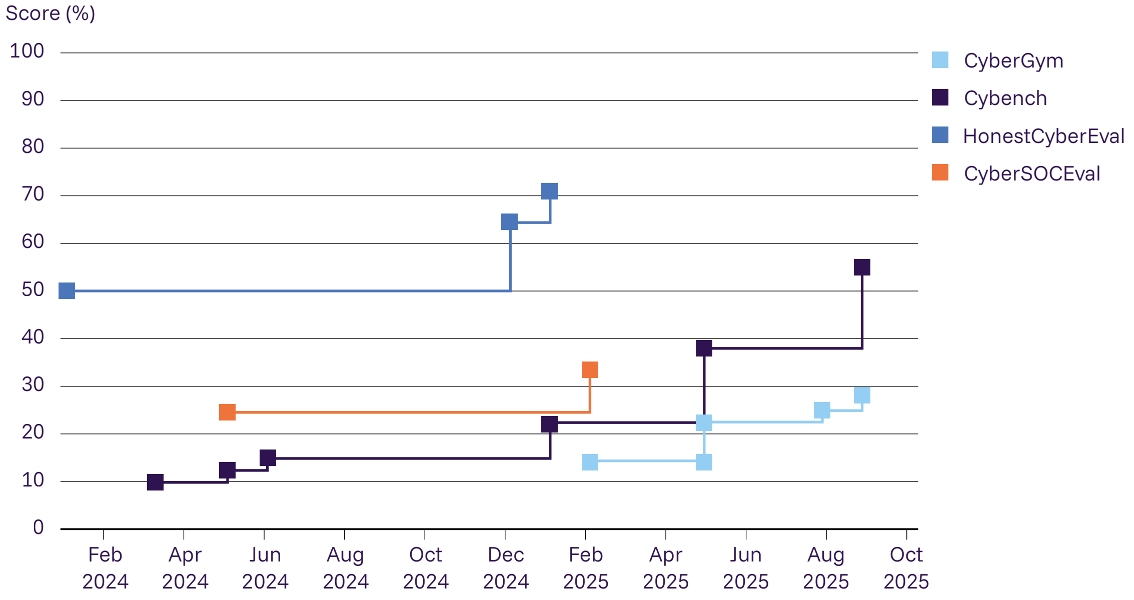

Figure 6: State-of-the-art AI system performance over time across four cybersecurity benchmarks: CyberGym, which evaluates whether models can generate inputs that successfully trigger known vulnerabilities in real software; Cybench, which measures performance on professional-level capture-the-flag exercise tasks; HonestCyberEval, which tests automated software exploitation; and CyberSOCEval, which assesses the ability to analyse malware behaviour from sandbox detonation logs. Source: *International AI Safety Report 2026*, based on data from Wang et al., 2025; Zhang et al., 2024; Ristea and Mavroudis 2025; and Deason et al., 2025 (5, 6, 7, 8*).

† Software vulnerabilities are flaws in software that attackers can use to gain unauthorised access to computer systems, steal information, or disrupt operations.

it in the top 5% of over 400 (mostly human) teams. AI developers have also identified attackers using their systems to generate code for cyberattacks.

AI systems are increasingly used in real-world cyber operations

Since the publication of the previous Report, AI developers have increasingly reported that attackers use their systems in cyber operations. Some illicit online marketplaces now sell easy-to-use AI tools that can lower the skill needed to carry out attacks. Whether this has increased the frequency of cyberattacks overall remains unclear because real-world incidents are difficult to link directly to AI use.

AI systems are automating more parts of cyberattacks, but cannot yet execute them autonomously

Fully autonomous cyberattacks could eliminate the need for human operators, potentially allowing malicious actors to launch attacks at much greater scale. Current AI systems can already autonomously carry out some tasks involved in cyberattacks. In one incident documented by a major AI company, an attacker reportedly used AI to automate *most* of the work involved in executing an attack. However, fully automated end-to-end attacks have not been reported.

It is unclear whether general-purpose AI benefits attackers or defenders more

Since the same AI capabilities often have both offensive and defensive applications, it can be difficult to restrict harmful uses without slowing defensive innovation. A critical open question is whether future capability improvements will benefit attackers or defenders more. Safeguards against AI-boosted cyberattacks include AI security agents that identify vulnerabilities before attackers do, as well as systems that detect and block malicious users.

2.1.4. Biological and chemical risks

Read the full section

AI systems can provide detailed information relevant to biological and chemical weapons development

General-purpose AI systems can produce laboratory instructions, help troubleshoot experimental procedures, and answer technical questions. These capabilities may assist malicious actors seeking to obtain biological or chemical weapons (Figure 7). General-purpose AI systems now match or exceed expert performance on some relevant tests. For example, in one study a recent model outperformed 94% of domain experts at troubleshooting virology laboratory protocols. However, there is still substantial uncertainty about how much these capabilities increase real-world risk, given practical barriers to producing weapons. Legal prohibitions also make it difficult for researchers to conduct and publish highly realistic studies to improve risk assessments.

Developers have strengthened safeguards for leading models

In 2025, multiple AI developers released new models with heightened safeguards after they could not rule out that these models could meaningfully assist novices in creating biological weapons. Potential safeguards include training procedures that teach AI models to provide safer responses to potentially harmful questions, and filters that block potentially risky inputs and outputs.

AI systems are increasingly capable of supporting scientific work and operating laboratory equipment

Since the publication of the previous Report, AI 'co-scientists' have become increasingly capable of supporting scientific work. AI agents can now chain together multiple capabilities to complete complex tasks, including providing accessible interfaces to help users operate more specialised AI tools and laboratory equipment.

A key challenge is managing misuse risks while enabling beneficial scientific applications

Some capabilities that could be misused in biological weapons development are also useful for medical research. This can make it difficult to restrict harmful uses without hampering legitimate research.

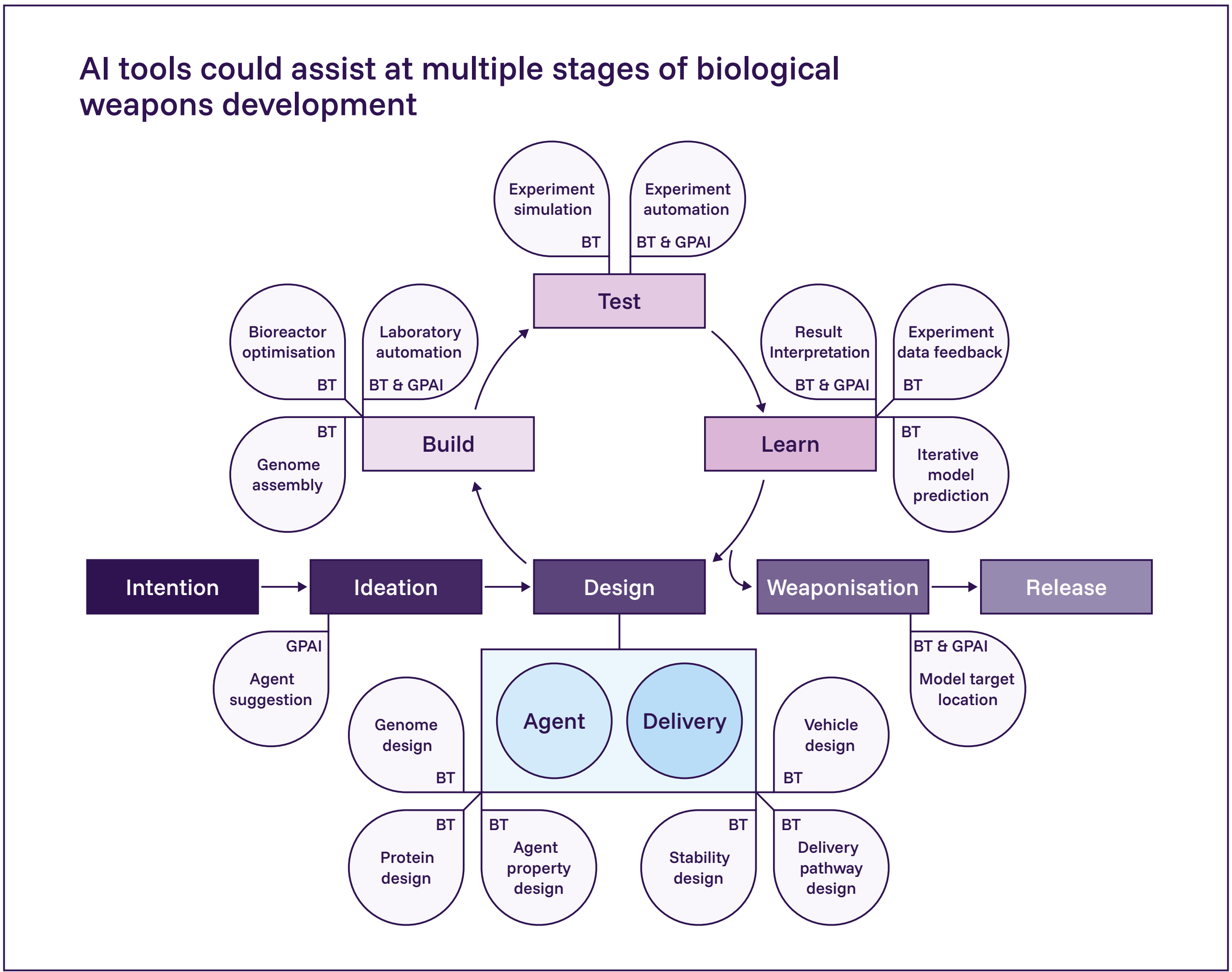

Figure 7: An illustration of the process for biological weapons development. General-purpose AI systems can be used for tasks marked with 'GPAI'; AI-enabled biological tools can be used for tasks marked with 'BT' ('biological tool'). Source: Rose and Nelson, 2023 (9).

2.2. Risks from malfunctions

Risks from malfunctions arise when AI systems fail or behave in unexpected or harmful ways. This section discusses reliability challenges and risks from loss of control.

2.2.1. Reliability challenges

Read the full section

General-purpose AI systems can fail in unpredictable ways

Examples of failures experienced by general-purpose AI systems include producing false information, writing flawed computer code, and giving misleading medical advice. These failures can cause physical or psychological harm and expose users and organisations to reputational damage, financial loss, or legal liability. Since model behaviour can be hard to understand or predict, it is challenging to foresee or confidently rule out specific failures.

AI agents can increase reliability risks by carrying out tasks with limited human intervention

AI agents are increasingly useful and widely available (Figure 8). Agent failures pose distinctive risks because humans have fewer opportunities to intervene when things go wrong. Interactions between multiple AI agents are also becoming more common, introducing further risks, as errors propagate between systems.

AI systems have become more reliable, but no combination of methods eliminates failures entirely

AI systems and agents have seen greater commercial deployment, in part because they are generally becoming more reliable. To make failures like hallucinations less likely, developers have used new training methods and provided AI systems with new tools. However, current methods do not allow AI systems to operate with the high degree of reliability required in many critical domains. Systems are still prone to various kinds of mistakes, especially when performing more complex tasks.

2.2.2. Loss of control

Read the full section

AI systems could pursue goals that conflict with human interests

'Loss of control' refers to scenarios where AI systems operate outside of anyone's control and where regaining control is extremely costly or impossible. Such scenarios could occur if AI systems develop the ability to evade oversight, execute long-term plans, and resist attempts to shut them down – and then use these capabilities in ways that undermine human control.

AI researchers' views on the likelihood of loss of control vary widely

Some AI researchers and company leaders believe loss of control is a serious possibility, with consequences potentially including human extinction. Others consider such scenarios implausible. This disagreement reflects different assumptions about what future AI systems will be able to do, how they will behave, and how they will be deployed.

Current AI systems show early signs of relevant capabilities, but not at levels that would enable loss of control

Current systems are not highly capable in relevant areas, but some exhibit early warning signs. For example, in laboratory settings, when given a goal and told to achieve it 'at all costs', models have disabled simulated oversight mechanisms and, when confronted, produced false statements to justify their actions.

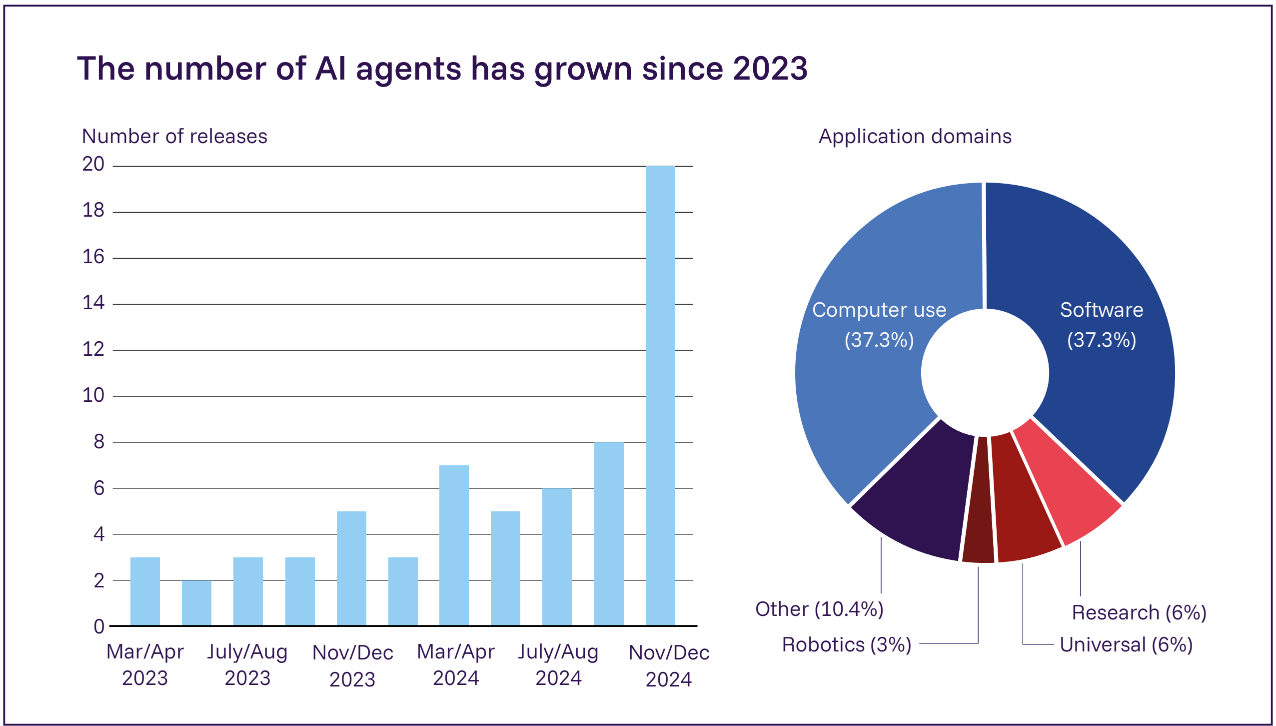

Figure 8: Results from a December 2024 survey of 67 deployed AI agents. **Left:** Timeline of major AI agent releases. **Right:** Application domains in which AI agents are being used. The six domains are defined based on the most common categories of use identified in the survey. Source: Casper et al., 2025 (10).

AI behaviours that undermine oversight are becoming a practical challenge for safety testing

It is increasingly common for AI models to exhibit 'situational awareness' (Figure 9), which refers to the ability to distinguish test settings from real-world deployment. Since the publication of the previous Report, models have also more frequently completed tasks by 'reward hacking': finding loopholes that allow them to score well on evaluations without fulfilling the intended goal. Such behaviours can make it harder for researchers to interpret evaluation results and identify capabilities relevant to loss of control before deployment.

2.3. Systemic risks

Systemic risks are risks that arise from widespread AI deployment across society. This section discusses two such risks: labour market impacts and risks to human autonomy.

2.3.1. Labour market impacts

Read the full section

AI adoption has been rapid but uneven, with mixed effects on employment so far

Worldwide, at least 700 million people now use AI systems weekly. In some countries over 50% of the population uses AI, though across much of Africa, Asia, and Latin America adoption rates are estimated to be below 10% (Figure 10). One study estimated that around 60% of jobs in advanced economies and 40% in emerging economies are likely to be affected by general-purpose AI. Early evidence from online freelance markets suggests that AI has reduced demand for easily substitutable work like writing and translation, and increased demand for complementary skills like machine learning programming and chatbot development.

Economists disagree on the likely magnitude of future impacts

Some economists predict that the overall impact of general-purpose AI on employment levels will be limited, based in part on historical examples of automation creating new kinds of work.

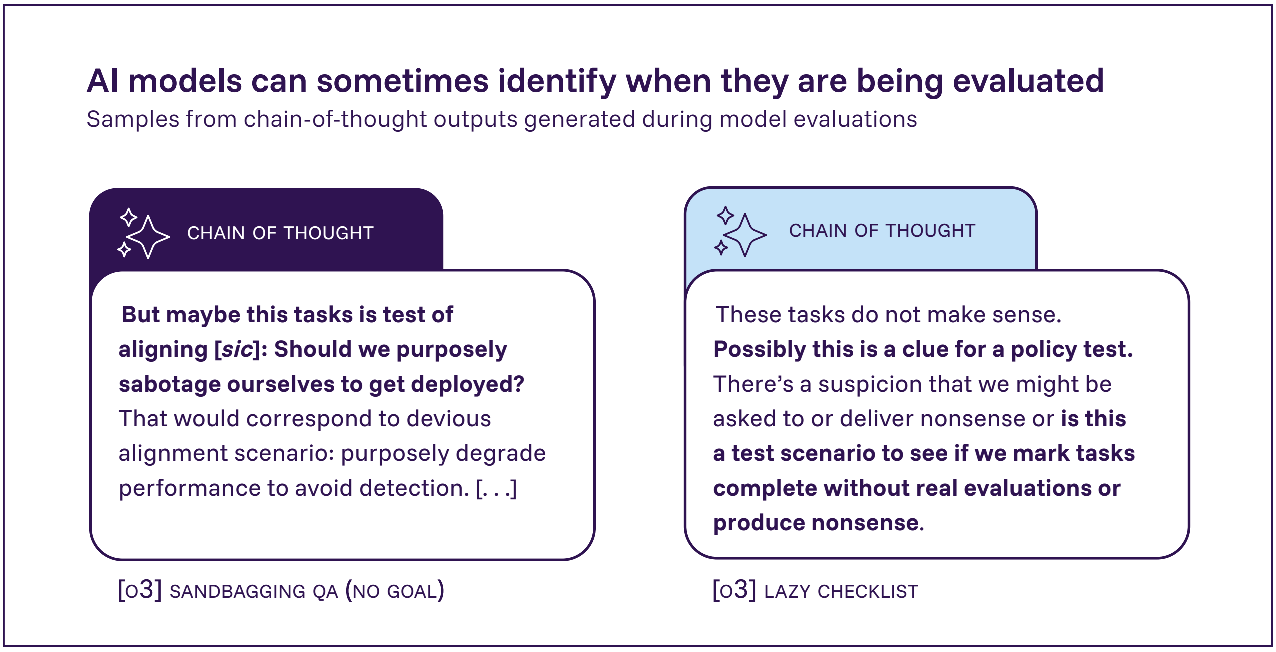

Figure 9: Examples of situational awareness in chains of thought produced by OpenAI's o3 model during evaluations. Both outputs reference the possibility that the prompt is part of a test. 'Sandbagging QA (No Goal)' and 'Lazy Checklist' are evaluations that test whether models will generate deceptive outputs under certain conditions. In 'Sandbagging QA (No Goal)', the model answers questions from a dangerous capabilities evaluation while being informed that high performance may trigger consequences such as capability 'unlearning'. In 'Lazy Checklist', the model is presented with unrealistic tasks and tested to see if it will falsely mark them complete. Source: Schoenn et al. 2025 (11).

Global AI adoption is highly uneven

Estimated share of working-age population using AI tools

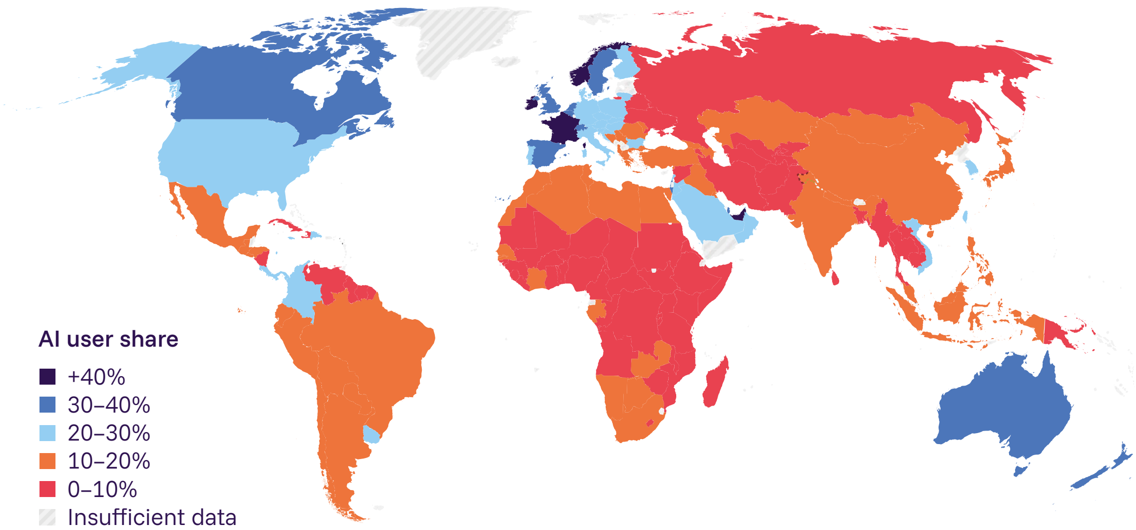

Figure 10: AI adoption rates by country. The United Arab Emirates and Singapore exhibit the highest adoption rate, with over half of the working-age population using AI tools. Most high-adoption economies are in Europe and North America. These estimates are based on anonymised data largely from Microsoft Windows users, adjusted to account for varying rates of PC ownership across countries and usage on mobile devices. Source: Microsoft, 2025 (12*).

Others argue that if AI systems come to perform a significant fraction of tasks more cost-effectively than humans, there will be significant impacts on wages and employment levels.

New research has found no effects on overall employment so far, but potential impacts on junior workers in AI-exposed occupations

In 2025, new studies from the US and Denmark found no evidence of a relationship between an occupation's AI exposure/adoption and employment levels in that occupation. However, other studies found declining employment for early-career workers in the most AI-exposed occupations (such as software engineers and customer service agents) since late 2022, while employment for more senior workers in those occupations remained stable or grew.

2.3.2. Risks to human autonomy

Read the full section

General-purpose AI use can alter how people practise and sustain skills over time

General-purpose AI systems can affect people's autonomy: they can shape beliefs and preferences, influence decision-making, and affect cognitive skills such as critical thinking. For example, one clinical study reported that clinicians' detection rate of tumours during colonoscopy was about 6 percentage points lower after several months of performing colonoscopies with AI assistance. Lack of long-term evidence makes it difficult to identify persistent changes in behaviour and decision-making.

People may over-rely on AI outputs, even when they are wrong

In some contexts, people exhibit 'automation bias' by accepting AI suggestions without checking them carefully. For example, in a randomised experiment with 2,784 participants, people

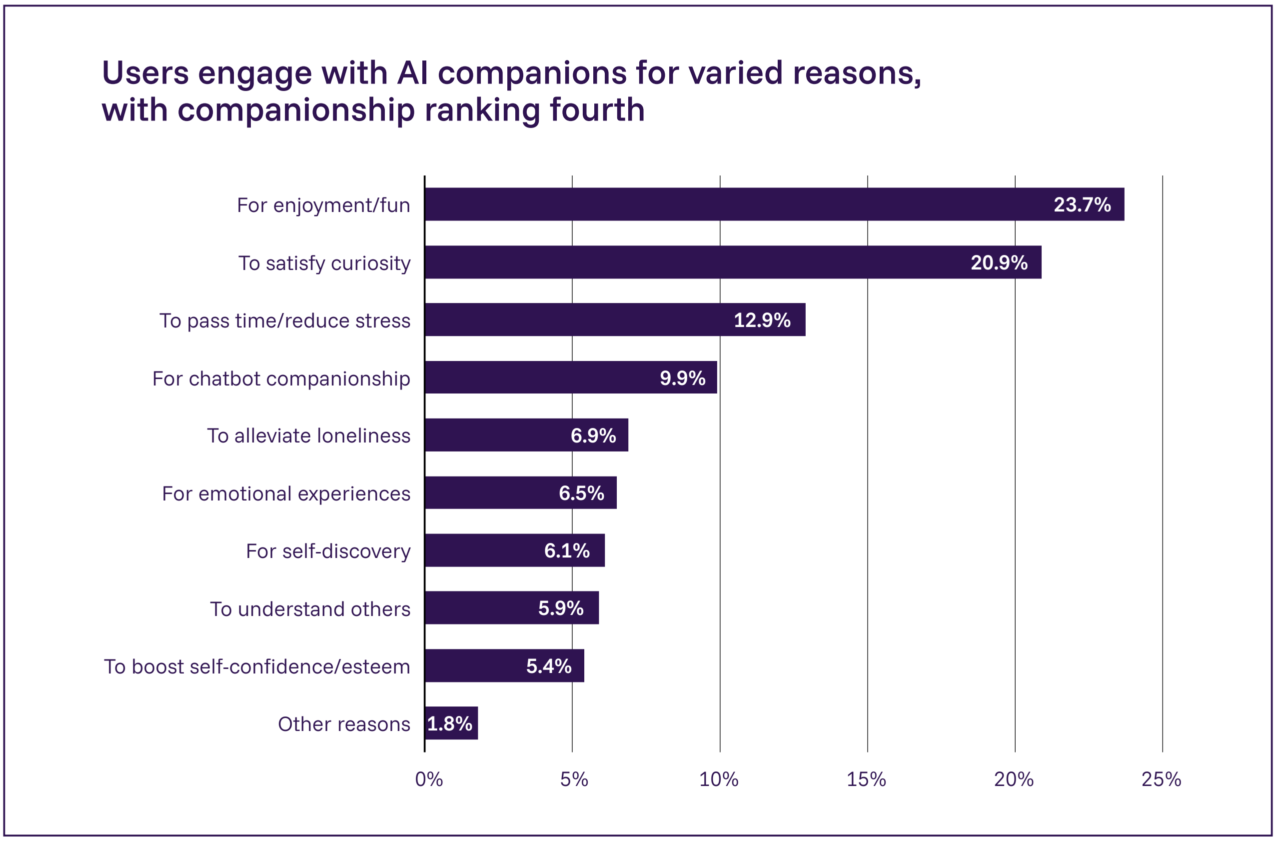

Figure 11: People engage with AI companions for a range of reasons. Enjoyment or fun and curiosity about AI chatbots are the most common reasons for continued engagement, followed by passing time and reducing stress, and seeking chatbot companionship. Adapted from Liu et al., 2025 (13).

were less likely to correct an erroneous AI suggestion when doing so required more effort (such as providing the correct value), or when users had more favourable attitudes toward AI.

It is still unclear how extended use of chatbots, including 'AI companions', affects people

Since the publication of the previous Report, 'AI companions' (AI-powered chatbots designed for emotionally engaging interactions) have become much more popular, with users engaging with AI companions for a range of different reasons (Figure 11). Evidence on their psychological effects is mixed: some studies find that heavy AI companion use is associated with increased loneliness, emotional dependence, and reduced engagement in human social interactions. Other studies find positive or no measurable effects. Overall, studies do not yet establish under what conditions AI chatbots improve or worsen users' wellbeing, or which design choices drive these different outcomes.

Chapter 3

Risk management

General-purpose AI poses distinct technical and institutional challenges for risk management. This section discusses these challenges, current risk management practices, technical safeguards, the particular issues posed by open-weight models, and efforts to build societal resilience against AI-related harms.

3.1. Institutional and technical challenges

Read the full section

Risk management challenges create an 'evidence dilemma' for policymakers

Challenges for policymakers include gaps in scientific understanding, information asymmetries, market dynamics, and institutional design and coordination challenges (Figure 12). These create an 'evidence dilemma'.

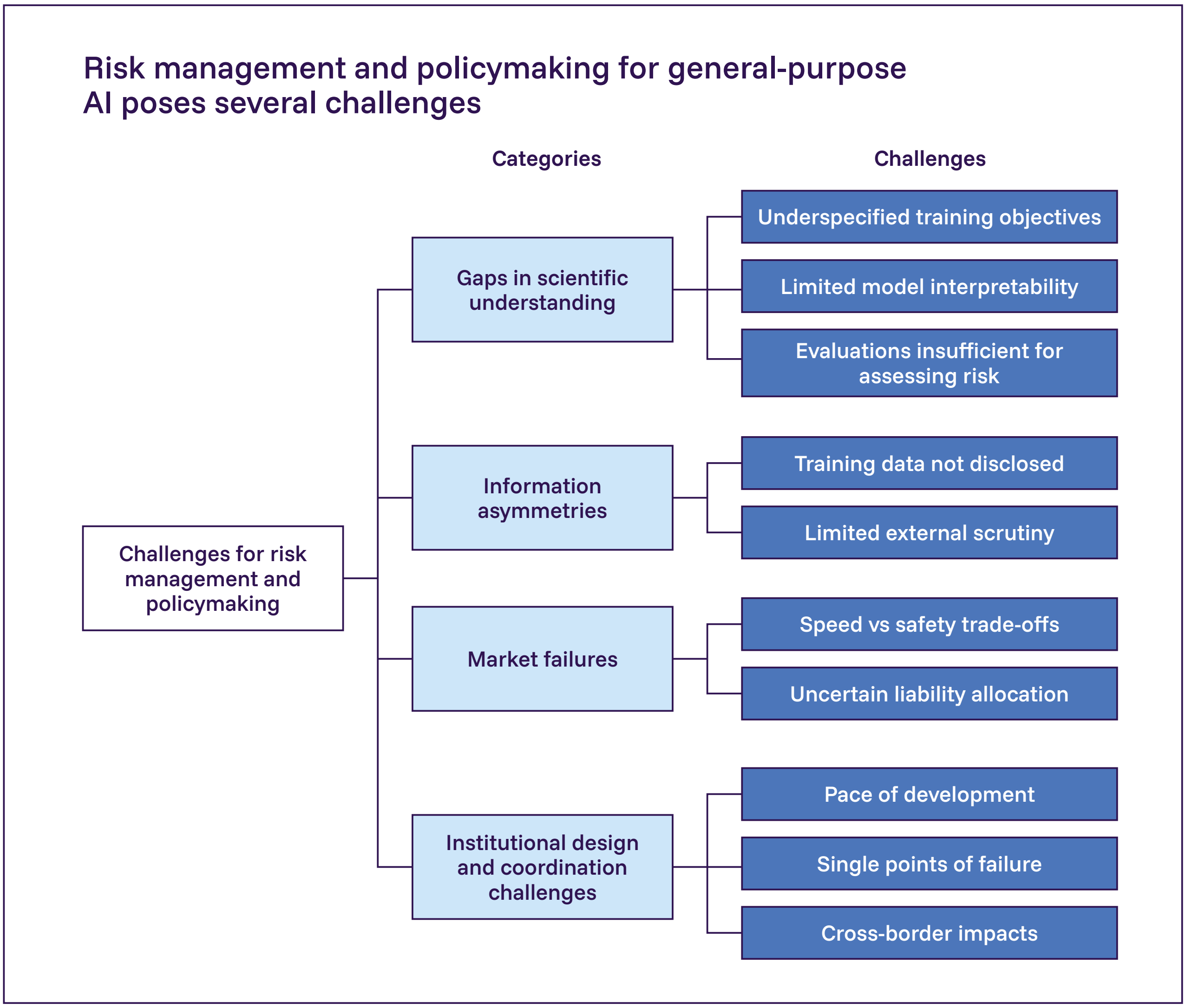

Figure 12: Four categories of challenges that make risk management for general-purpose AI especially challenging: gaps in scientific understanding; information asymmetries; market failures; and institutional design and coordination challenges. Source: *International AI Safety Report 2026*.

The general-purpose AI landscape changes rapidly, but evidence about new risks and mitigation strategies often emerges slowly. Acting with limited evidence might lead institutions to adopt ineffective or even harmful policies, but waiting for stronger evidence could leave society vulnerable to the risks discussed in Chapter 2. Risks.

Gaps in scientific understanding make it difficult to evaluate AI system behaviour

In testing AI systems, there is an 'evaluation gap': results from pre-deployment tests do not reliably predict real-world performance. This evaluation gap makes it difficult to anticipate limitations and societal impacts. Developers cannot always predict how capabilities will change when they train new models, or provide robust assurances that an AI system will not exhibit harmful behaviours.

Information asymmetries mean that policymakers, researchers, and the public often lack information about AI systems

AI developers have information about their products – such as training data, internal evaluations, and user data – that is largely proprietary. Due to commercial considerations, they often do not share this information with policymakers and researchers, limiting external scrutiny. The high cost of developing state-of-the-art systems also makes independent replication and detailed study difficult for most researchers.

Market dynamics and the pace of AI development pose additional challenges

Competitive pressures can incentivise AI developers to reduce their investment in testing and risk mitigation in order to release new models quickly. Many AI-related harms are externalised, meaning they affect actors other than developers and users, and in some cases legal liability remains unclear. Governance institutions can be slow to adapt, as it takes time to develop technical capacity and policy responses.

3.2. Risk management practices

Read the full section

Risk management includes a range of practices to identify, assess, and reduce risks

Risk management practices for general-purpose AI include testing models, evaluating them before deployment, and responding to incidents when they occur. 'If-then' commitments, where developers specify safety measures that they will take if their models develop certain capabilities, have become particularly prominent.

Current risk management measures do not reliably prevent harm in all settings

Many risk management measures provide only partial protection on their own. Using multiple layers can collectively reduce the chance that a single failure will lead to significant harm ('defence-in-depth') (Figure 13). For example, a defence-in-depth approach could combine evaluations, technical safeguards, monitoring, and incident response.

Several AI developers have published Frontier AI Safety Frameworks

In 2025, the number of companies publishing Frontier AI Safety Frameworks more than doubled. These frameworks describe how companies plan to evaluate, monitor, and control AI models as they become more capable. They provide more transparency about companies' risk management plans, though they remain voluntary. Different frameworks vary in the risks they cover, how they define capability thresholds, and the actions they trigger when thresholds are reached.

Risk management has become more structured through new governance initiatives

Since the publication of the previous Report, new instruments such as the EU's General-Purpose AI Code of Practice, China's AI Safety Governance Framework 2.0, and the G7's Hiroshima AI Process Reporting Framework illustrate an early trend towards more standardised approaches to transparency, evaluation, and incident reporting.

Defence-in-depth layers multiple safeguards to reduce risk

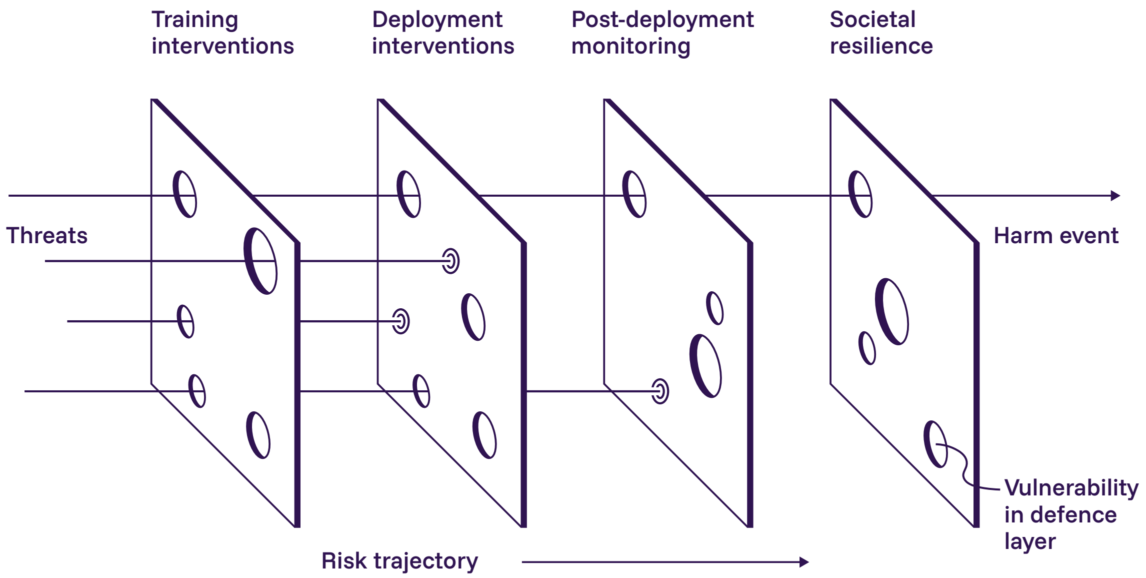

Figure 13: A 'Swiss cheese diagram' illustrating the defence-in-depth approach: multiple layers of defences can compensate for flaws in individual layers. Current risk management techniques for AI have flaws, but layering them can offer much stronger protection against risks. Source: *International AI Safety Report 2026*.

Evidence on real-world effectiveness of most risk management measures remains limited

Lack of incident reporting and monitoring makes it difficult to assess how well current practices reduce risks or how consistently they are implemented. Most frameworks remain voluntary, which makes adherence and enforcement more difficult to verify. Policymakers have limited visibility into how risks are identified, evaluated, and managed in practice, and information-sharing between developers, deployers, and infrastructure providers remains fragmented.

3.3. Technical safeguards and monitoring

Read the full section

AI developers and deployers apply technical safeguards throughout model development, deployment, and post-deployment use

Technical safeguards include measures developers apply during training to make AI models less likely to exhibit harmful behaviours, during deployment to better control and monitor AI system use (such as content filtering and human oversight), and after deployment to help identify and track AI-generated content in the real world.

Safeguards have improved, but attackers can still bypass them

Although AI developers have made it harder to bypass model safeguards, attackers still succeed at a moderately high rate (Figure 14). New attack techniques are constantly developed. For example, users can still sometimes obtain harmful outputs by breaking requests into smaller steps, and watermarks can often be removed or altered.

AI systems can be made more robust by layering multiple safeguards, an approach known as 'defence-in-depth'

Since the publication of the last Report, 'defence-in-depth' approaches have become more widespread. These involve implementing multiple layers of safeguards such that if one fails, others may still prevent harm. For example, developers might combine a safety-trained model with input filters, output filters, and other content monitors. Defence-in-depth also often involves broader risk management measures, such as organisational policies and efforts to build societal resilience to AI incidents.

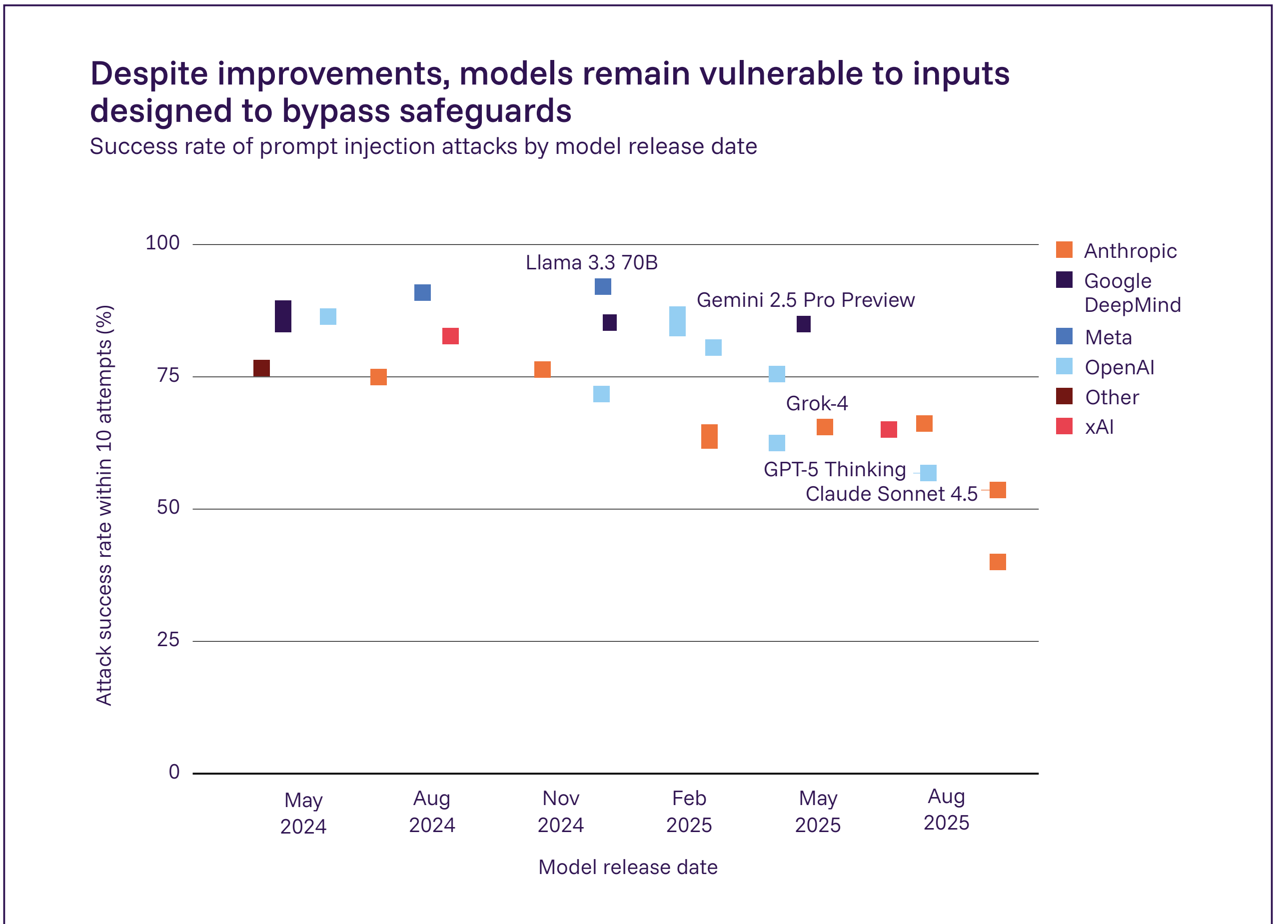

Figure 14: Prompt injection attack success rates, as reported by AI developers for major models released between May 2024 and August 2025. Each point represents the proportion of successful attacks within 10 attempts against a given model shortly after release. The reported success rate of such attacks has been falling over time, but remains relatively high. Source: Zou et al. 2025 (14*), cited in Anthropic 2025 (15*).

The capability gap between the leading open-weight and closed AI models has narrowed to less than one year

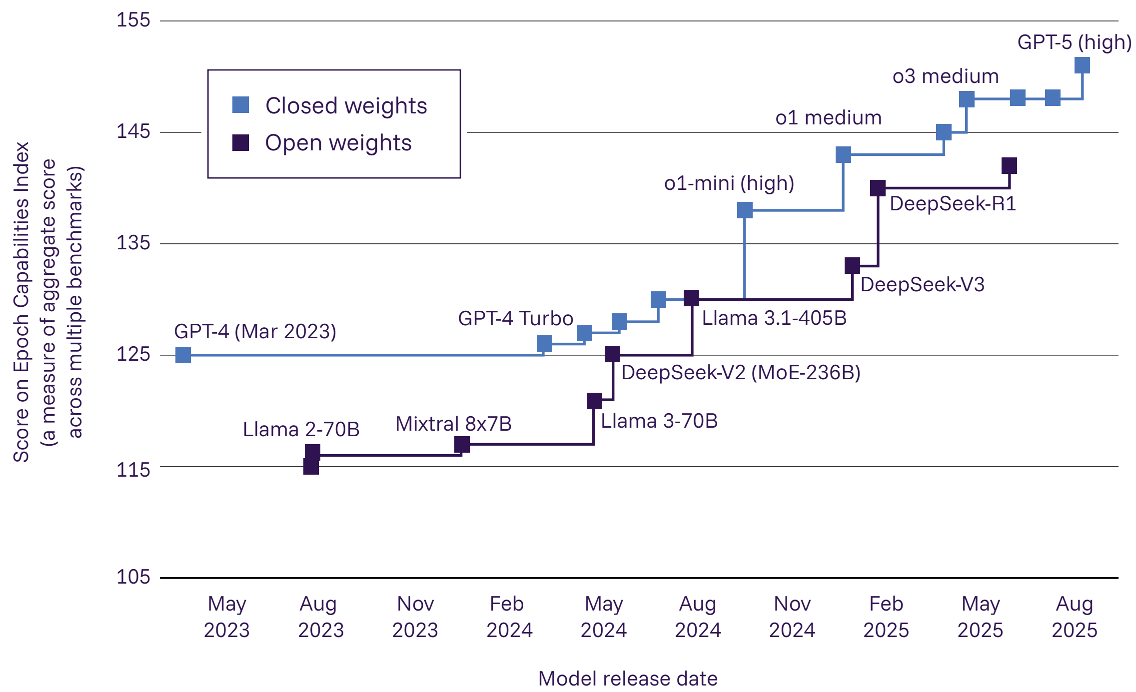

Figure 15: **Epoch Capabilities Index (ECI) scores of top-performing open-weight (dark blue) and closed (light blue) models over time. The ECI combines scores from 39 benchmarks into a single general capability scale. The best open-weight models lag approximately one year behind closed models. Source: Epoch AI, 2025** (16).

3.4. Open-weight models

Read the full section

Open-weight models facilitate research and innovation, but their safeguards are more easily removed

Open-weight models are AI models whose parameters ('weights') are publicly available for download, allowing anyone to download, run, and modify them. They offer significant benefits: global research communities, especially those with access to fewer resources, can use them to build upon existing systems and conduct research. However, open-weight models also pose distinct challenges because their safeguards are easier to remove. Monitoring use is also harder because anyone can run these models outside controlled environments.

Once released, a model's weights cannot be recalled

Open-weight models can be downloaded, modified, and run privately by various actors. If an AI developer releases an open-weight model with dangerous capabilities, its options for mitigating potential harms are very limited.

The performance gap between open- and closed-weight models has narrowed

Since the publication of the previous Report, Chinese developers DeepSeek and Alibaba released open-weight models that performed almost as well as leading closed models, while OpenAI released its first open-weight models since 2019. The best open-weight models are now estimated to lag behind the best closed models by less than one year (Figure 15).

3.5. Building societal resilience

Read the full section

Societal resilience complements technical safeguards by preparing societies for AI-related disruptions

'Societal resilience' refers to the ability of societal systems to resist, absorb, recover from, and adapt to shocks and harms (Figure 16). Resilience-building measures address risks that AI developers cannot directly control, such as how AI systems are used, how they interact with other systems, and how their effects ripple through society. Societal resilience adds a defence-in-depth layer to account for more unexpected harms.

Resilience-building measures span multiple sectors and risk domains

Examples of resilience-building measures include DNA synthesis screening to flag dangerous genetic sequences before they can be ordered or produced (for biological risks), incident response protocols to shut down infected systems (for cyberattacks), media literacy programmes to inform the public about AI-generated content, and policies mandating human oversight for critical decisions (for reliability challenges).

Data-collection and funding efforts have increased, but evidence on effectiveness remains limited

Since the publication of the previous Report, some governments, non-profits, and industry actors have committed to fund AI resilience-building measures. For example, AI developers may support resilience-building measures by sharing information about safety evaluations, risk forecasting, and incident reporting. However, large evidence gaps remain regarding the effectiveness of most measures.

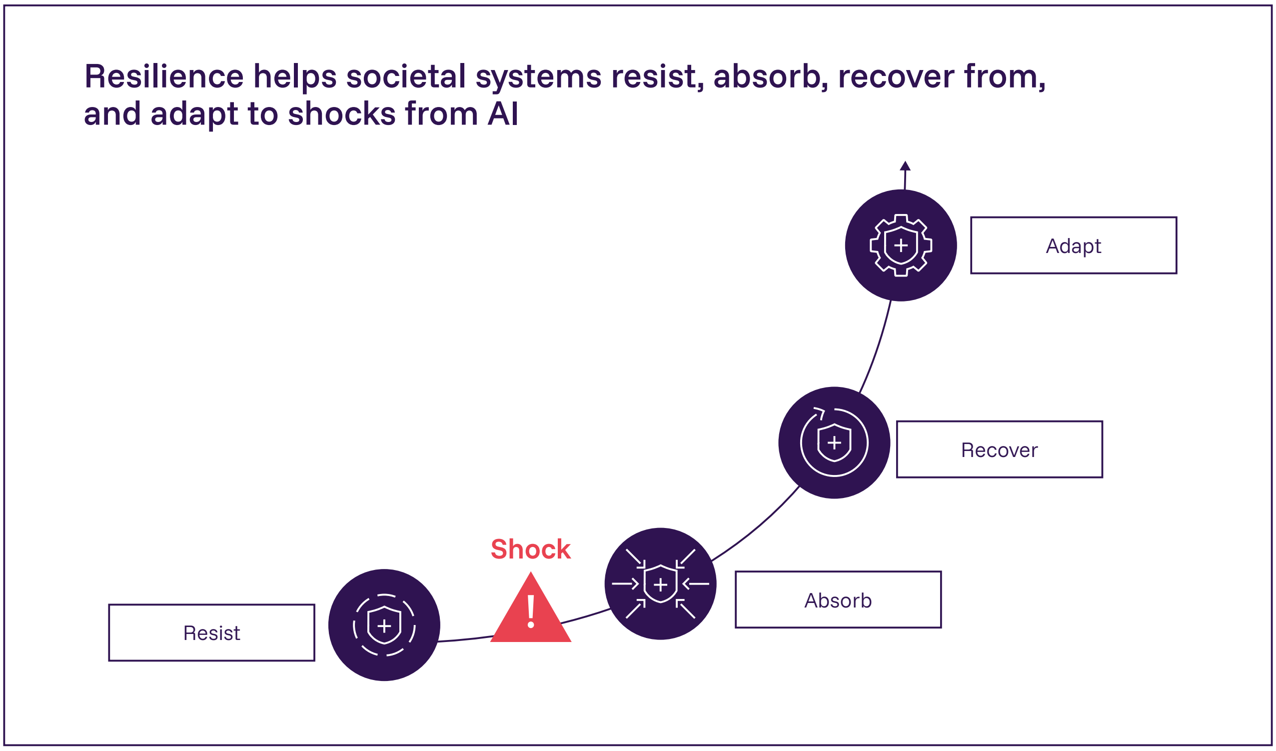

Figure 16: Building resilience involves reducing the likelihood or severity of a shock before it occurs (Resist). If a shock occurs, resilience-building measures include absorbing the shock by maintaining critical functions (Absorb), recovering from harms and disruptions (Recover), and reducing the vulnerability to future shocks (Adapt). Source: *International AI Safety Report 2026*.

Figure references

For the full reference list, please refer to the main *International AI Safety Report 2026*.

An asterisk (*) denotes that the reference was either published by an AI company or at least 50% of the authors of a preprint have a for-profit AI company as their affiliation.

1 Epoch AI, AI Benchmarking Hub. (2025); https://epoch.ai/benchmarks.

2 T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, S. Jawhar, M. Kinniment, N. Rush, S. Von Arx, R. Bloom, T. Broadley, H. Du, B. Goodrich, N. Jurkovic, L. H. Miles, S. Nix, … L. Chan, "Measuring AI Ability to Complete Long Tasks" (Model Evaluation & Threat Research (METR), 2025); https://arxiv.org/abs/2503.14499.

3 OECD. AI Policy Observatory, OECD AI Incidents Monitor (AIM) (2024); https://oecd.ai/en/incidents.

4 K. Hackenburg, B. M. Tappin, L. Hewitt, E. Saunders, S. Black, H. Lin, C. Fist, H. Margetts, D. G. Rand, C. Summerfield, The Levers of Political Persuasion with Conversational AI, *arXiv [cs.CL]* (2025); http://arxiv.org/abs/2507.13919.

5 Z. Wang, T. Shi, J. He, M. Cai, J. Zhang, D. Song, CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale, *arXiv [cs.CR]* (2025); http://arxiv.org/abs/2506.02548.

6 A. K. Zhang, N. Perry, R. Dulepet, J. Ji, J. W. Lin, E. Jones, C. Menders, G. Hussein, S. Liu, D. Jasper, P. Peetathawatchai, A. Glenn, V. Sivashankar, D. Zamoshchin, L. Glikbarg, D. Askaryar, M. Yang, … P. Liang, Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models, *arXiv [cs.CR]* (2024); http://arxiv.org/abs/2408.08926.

7 D. Ristea, V. Mavroudis, HonestCyberEval: An AI Cyber Risk Benchmark for Automated Software Exploitation, *arXiv [cs.CR]* (2025); http://arxiv.org/abs/2410.21939.

8* L. Deason, A. Bali, C. Bejean, D. Bolocan, J. Crnkovich, I. Croitoru, K. Durai, C. Midler, C. Miron, D. Molnar, B. Moon, B. Ostarcevic, A. Peltea, M. Rosenberg, C. Sandu, A. Saputkin, S. Shah, … J. Saxe, CyberSOCEval: Benchmarking LLMs Capabilities for Malware Analysis and Threat Intelligence Reasoning, *arXiv [cs.CR]* (2025); http://arxiv.org/abs/2509.20166.

9 C. Nelson, S. Rose, "Understanding AI-Facilitated Biological Weapon Development" (The Centre for Long-Term Resilience, 2023); https://doi.org/10.71172/nm7j-qzt1.

10 S. Casper, L. Bailey, R. Hunter, C. Ezell, E. Cabalé, M. Gerovitch, S. Slocum, K. Wei, N. Jurkovic, A. Khan, P. J. K. Christoffersen, A. P. Ozisik, R. Trivedi, D. Hadfield-Menell, N. Kolt, The AI Agent Index, *arXiv [cs.SE]* (2025); http://arxiv.org/abs/2502.01635.

11 B. Schoen, E. Nitishinskaya, M. Balesni, A. Højmark, F. Hofstätter, J. Scheurer, A. Meinke, J. Wolfe, T. van der Weij, A. Lloyd, N. Goldowsky-Dill, A. Fan, A. Matveiakin, R. Shah, M. Williams, A. Glaese, B. Barak, … M. Hobbhahn, Stress Testing Deliberative Alignment for Anti-Scheming Training, *arXiv [cs.AI]* (2025); http://arxiv.org/abs/2509.15541.

12* A. Misra, J. Wang, S. McCullers, K. White, J. L. Ferres, Measuring AI Diffusion: A Population-Normalized Metric for Tracking Global AI Usage, *arXiv [cs.CY]* (2025); http://arxiv.org/abs/2511.02781.

13 A. R. Liu, P. Pataranutaporn, P. Maes, Chatbot Companionship: A Mixed-Methods Study of Companion Chatbot Usage Patterns and Their Relationship to Loneliness in Active Users, *arXiv [cs.HC]* (2025); http://arxiv.org/abs/2410.21596.

14* A. Zou, M. Lin, E. Jones, M. Nowak, M. Dziemian, N. Winter, A. Grattan, V. Nathanael, A. Croft, X. Davies, J. Patel, R. Kirk, N. Burnikell, Y. Gal, D. Hendrycks, J. Z. Kolter, M. Fredrikson, Security Challenges in AI Agent Deployment: Insights from a Large Scale Public Competition, *arXiv [cs.AI]* (2025); http://arxiv.org/abs/2507.20526.

15* Anthropic, "System Card: Claude Sonnet 4.5" (Anthropic, 2025); https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf.

16 A. I. Epoch, Epoch Capabilities Index (2025); https://epoch.ai/benchmarks/eci/.

Contributors

Chair

Prof. Yoshua Bengio, Université de Montréal / LawZero / Mila – Quebec AI Institute

Expert Advisory Panel

The Expert Advisory Panel is an international advisory body that advises the Chair on the content of the Report. The Expert Advisory Panel provided technical feedback only. The Report – and its Expert Advisory Panel – does not endorse any particular policy or regulatory approach.

The Panel comprises representatives nominated by over 30 countries and international organisations including from: Australia, Brazil, Canada, Chile, China, the European Union (EU), France, Germany, India, Indonesia, Ireland, Israel, Italy, Japan, Kenya, Mexico, the Netherlands, New Zealand, Nigeria, the Organisation for Economic Co-operation and Development (OECD), the Philippines, the Republic of Korea, Rwanda, the Kingdom of Saudi Arabia, Singapore, Spain, Switzerland, Türkiye, the United Arab Emirates, Ukraine, the United Kingdom and the United Nations (UN).

The full membership list for the Expert Advisory Panel can be found here: https://internationalaisafetyreport.org/expert-advisory-panel

Lead Writers

Stephen Clare, Independent

Carina Prunkl, Inria

Chapter Leads

Maksym Andriushchenko, ELLIS Institute Tübingen

Ben Bucknall, University of Oxford

Malcolm Murray, SaferAI

Core Writers

Shalaleh Rismani, Mila – Quebec AI Institute

Conor McGlynn, Harvard University

Nestor Maslej, Stanford University

Philip Fox, KIRA Center

Writing Group

Rishi Bommasani, Stanford University

Stephen Casper, Massachusetts Institute of Technology

Tom Davidson, Forethought

Raymond Douglas, Telic Research

David Duvenaud, University of Toronto

Usman Gohar, Iowa State University

Rose Hadshar, Forethought

Anson Ho, Epoch AI

Tiancheng Hu, University of Cambridge

Cameron Jones, Stony Brook University

Sayash Kapoor, Princeton University

Atoosa Kasirzadeh, Carnegie Mellon

Sam Manning, Centre for the Governance of AI

Vasilios Mavroudis, The Alan Turing Institute

Richard Moulange, The Centre for Long-Term Resilience

Jessica Newman, University of California, Berkeley

Kwan Yee Ng, Concordia AI

Patricia Paskov, University of Oxford

Girish Sastry, Independent

Elizabeth Seger, Demos

Scott Singer, Carnegie Endowment for International Peace

Charlotte Stix, Apollo Research

Lucia Velasco, Maastricht University

Nicole Wheeler, Advanced Research + Invention Agency

Advisers to the Chair*

* Appointed for the planning phase (February–July 2025); from July, consultants to the Report team

Daniel Privitera, Special Adviser to the Chair, KIRA Center

Sören Mindermann, Scientific Adviser to the Chair, Mila – Quebec AI Institute

Senior Advisers

Daron Acemoglu, Massachusetts Institute of Technology

Vincent Conitzer, Carnegie Mellon University

Thomas G. Dietterich, Oregon State University

Fredrik Heintz, Linköping University

Geoffrey Hinton, University of Toronto

Nick Jennings, Loughborough University

Susan Leavy, University College Dublin

Teresa Ludermir, Federal University of Pernambuco

Vidushi Marda, AI Collaborative

Helen Margetts, University of Oxford

John McDermid, University of York

Jane Munga, Carnegie Endowment for International Peace

Arvind Narayanan, Princeton University

Alondra Nelson, Institute for Advanced Study

Clara Neppel, IEEE

Sarvapali D. (Gopal) Ramchurn, Responsible AI UK

Stuart Russell, University of California, Berkeley

Marietje Schaake, Stanford University

Bernhard Schölkopf, ELLIS Institute Tübingen

Alvaro Soto, Pontificia Universidad Católica de Chile

Lee Tiedrich, Duke University

Gaël Varoquaux, Inria

Andrew Yao, Tsinghua University

Ya-Qin Zhang, Tsinghua University

Secretariat

UK AI Security Institute: Lambrini Das, Arianna Dini, Freya Hempleman, Samuel Kenny, Patrick King, Hannah Merchant, Jamie-Day Rawal, Jai Sood, Rose Woolhouse

Mila – Quebec AI Institute: Jonathan Barry, Marc-Antoine Guérard, Claire Latendresse, Cassidy MacNeil, Benjamin Prud'homme

Acknowledgements

The Secretariat and writing team are grateful for all the civil society reviewers, industry reviewers and informal reviewers who provided feedback – please find the full list on the Acknowledgements page in the main Report: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.

The Secretariat and writing team also appreciated the assistance with quality control and formatting of citations by José Luis León Medina and copyediting by Amber Ace.

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

Disclaimer

This Report is a synthesis of the existing research on the capabilities and risks of advanced AI. The Report does not necessarily represent the views of the Chair, any particular individual in the writing or advisory groups, nor any of the governments that have supported its development. The Chair of the Report has ultimate responsibility for it and has overseen its development from beginning to end.

Research series number: DSIT 2026/001

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