

Second International AI Safety Report finds agents doing 30-minute coding tasks

Source copy kept by wheresthe.ai for the story “Second International AI Safety Report finds agents doing 30-minute coding tasks” (<https://wheresthe.ai/d/intl-international-ai-safety-report-2026-02/>).

Source: <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 5f393e4f9cb4b4b06bd4f0c54900638c0ff14561b5a863a03f9c863aa5fe17db

Documents the source offers: <https://www.dni.gov/files/ODNI/documents/assessments/ATA-2025-Unclassified-Report.pdf>;
<https://www.eeas.europa.eu/sites/default/files/documents/2025/EEAS-3nd-ThreatReport-March-2025-05-Digital-HD.pdf>; <https://businessdefense.gov/ibr/pae/docs/AI-RFI-Summary-Report.pdf>

Open Table of contents

Table of contents

Contributors

Chair

Prof. Yoshua Bengio , Université de Montreal / 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 organisation 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

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

Civil Society and Industry Reviewers

Civil Society

Ada Lovelace Institute, African Centre for Technology Studies, AI Forum New Zealand / Te Kāhui Atamai Iahiko o Aotearoa, AI Safety Asia, Stichting Algorithm Audit, Carnegie Endowment for International Peace, Center for Law and Innovation / Certa Foundation, Centre for the Governance of AI, Chief Justice Meir Shamgar Center for Digital Law and Innovation, Digital Futures Lab, EON Institute, Equiano Institute, Good Ancestors Policy, Gradient Institute, Institute for Law & AI, Interface, Israel Democracy Institute, Mozilla Foundation, NASSCOM, Old Ways New, RAND, Royal Society, SaferAI, Swiss Academy of Engineering Sciences, The Centre for Long-Term Resilience, The Alan Turing Institute, The Ethics Centre, The Future Society, The HumAIne Foundation, Türkiye Artificial Intelligence Policies Association

Industry

Advai, Anthropic, Cohere, Deloitte, Digital Umuganda, Domyne, G42, Google DeepMind, Harmony Intelligence, Hugging Face, HumAIn, IBM, LG AI Research, Meta, Microsoft, Naver, OpenAI, Qhala

Informal reviewers

Markus Anderljung, David Autor, Mariette Awad, Jamie Bernardi, Stella Biderman, Asher Brass, Ben Brooks, Miles Brundage, Kevin Bryan, Rafael Calvo, Siméon Campos, Carmen Carlan, Micah Carroll, Alan Chan, Jackie Cheung, Josh Collyer, Elena Cryst, Tino Cuéllar, Allan Dafoe, Jean-Stanislas Denain, Fernando Diaz, Roel Dobbe, Seth Donoughe, Seb Elmes, Izzy Gainsbury, Ben Garfinkel, Adam Gleave, Jasper Götting, Kobi Hackenburg, Lewis Hammond, David Evan Harris, Dan Hendrycks, José Hernández-Orallo, Luke Hewitt, Marius Hobbhahn, Manoel Horta Ribeiro, Abigail Jacobs, Ari Kagan, Daniel Kang, Anton Korinek, Michal Kosinski,

Gretchen Krueger, Dan Lahav, Anton Leicht, Vera Liao, Eli Lifland, Matthijs Maas, James Manyika, Simon Mylius, AJung Moon, Seán Ó hÉigeartaigh, Tamara Paris, Raymond Perrault, Siva Reddy, Luca Righetti, Jon Roozenbeek, Max Roser, Anders Sandberg, Leo Schwinn, Jaime Sevilla, Theodora Skeadas, Chandler Smith, Tobin South, Jonathan Spring, Merlin Stein, David Stillwell, Daniel Susser, Helen Toner, Sander van der Linden, Kush Varshney, Jess Whittlestone, Kai-Cheng Yang

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

Copyright and disclaimer

© Crown owned 2026

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/> or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or email: .

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned.

Any enquiries regarding this publication should be sent to: .

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

Forewords

A new scientific assessment of a fast-moving technology

Professor Yoshua Bengio – Université de Montréal / LawZero / Mila – Quebec AI Institute & Chair

This is the second International AI Safety Report , which builds on the mandate by world leaders at the 2023 AI Safety Summit at Bletchley Park to produce an evidence base to inform critical decisions about general-purpose artificial intelligence (AI).

This year, we have introduced several changes to make this Report even more useful and accessible.

First, to help policymakers better understand the range of potential outcomes despite the uncertainty involved, we have drawn upon new research conducted by the Organisation for Economic Co-operation and Development (OECD) and Forecasting Research Institute to present more specific scenarios and forecasts.

Second, following extensive consultation, we have narrowed the scope to focus on ‘emerging risks’: risks that arise at the frontier of AI capabilities. Given high uncertainty in this domain, the rigorous analysis the Report provides can be especially valuable. A narrower scope also ensures this Report complements other efforts, including the United Nations’ Independent International Scientific Panel on AI.

Of course, some things have not changed.

This remains the most rigorous assessment of AI capabilities, risks, and risk management available. Its development involved contributions from over 100 experts, including the guidance of experts nominated by over 30 countries and intergovernmental organisations.

The Report’s fundamental goal is also the same: to advance a shared understanding of how AI capabilities are evolving, risks associated with these advances, and what techniques exist to mitigate those risks.

The pace of AI progress raises daunting challenges. However, working with the many experts that produced this Report has left me hopeful. I am immensely grateful for the enormous efforts of all contributors – we are making progress towards understanding these risks.

With this Report, we hope to improve our collective understanding of what may be the most significant

technological transformation of our time.

Building a secure future for AI through international cooperation

Kanishka Narayan MP – Minister for AI and Online Safety, Department for Science, Innovation and Technology, UK Government

AI continues to redefine the possibilities before us – transforming economies, revitalising public services, and rapidly accelerating scientific advancement. This pace of progress demands an up-to-date, shared understanding of AI capabilities. This effort will build trust, enable adoption and pave the way for AI to deliver prosperity for all.

The 2026 International AI Safety Report is the result of strong collaboration across countries, organisations, civil society and industry partners – working together to produce robust, evidence-based analysis. The Report provides an essential tool for policymakers and world leaders to help navigate this challenging and fast-moving landscape.

The United Kingdom remains committed to strengthening international partnerships, scientific collaboration, and institutions that drive innovative AI research forward, including the AI Security Institute. Following the success of the landmark Summits hosted in Bletchley Park (November 2023), Seoul (May 2024) and Paris (February 2025), I am especially looking forward to the India AI Impact Summit – where this Report will be showcased – to ensure AI is shaped for humanity, inclusive growth and a sustainable future.

I am delighted to present this Report and thank Yoshua Bengio, the writing team, and all contributors for their dedication to this initiative. Together – through shared responsibility and international cooperation – we can forge a path where AI delivers security, opportunity and growth for every nation and every citizen.

Enabling equitable access to AI for all

Ashwini Vaishnaw – Minister of Railways, Information & Broadcasting and Electronics & Information Technology , Government of India

The second International AI Safety Report builds on the mandate of the 2023 AI Safety Summit at Bletchley Park. It aims at developing a shared, science-based understanding of advanced AI capabilities and risks.

This edition focuses on rapidly evolving general-purpose AI systems, including language, vision and agentic models. It also reviews associated challenges, including wider impacts on labour markets, human autonomy and concentration of power.

As AI systems grow more capable, safety and security remain critical priorities. The Report highlights practical approaches of model evaluations, dangerous capability thresholds and ‘if-then’ safety commitments to reduce high-impact failures.

Our global risk management frameworks are still immature, with limited quantitative benchmarks and significant evidence gaps. These gaps must be addressed alongside innovation.

For India and the Global South, AI safety is closely tied to inclusion, safety and institutional readiness. Responsible openness of AI models, fair access to compute and data, and international cooperation are essential too.

As host of the 2026 India AI Impact Summit, India has a key role in shaping global AI safety efforts. The Report is intended to help policymakers, researchers, industry and civil society shape national strategies.

About this Report

This is the second edition of the International AI Safety Report . The series was created following the 2023 AI Safety Summit at Bletchley Park to support an internationally shared scientific understanding of the capabilities and risks associated with advanced AI systems. A diverse group of over 100 Artificial Intelligence (AI) experts guided its development, including an international Expert Advisory Panel with nominees from over 30 countries and international organisations, including the Organisation for Economic Co-operation and Development (OECD), the European Union (EU), and the United Nations (UN).

Scope, focus, and independence

Scope: This Report concerns ‘general-purpose AI’: AI models and systems capable of performing a wide variety of tasks across different contexts. These models and systems perform tasks like generating text, images, audio, or

other forms of data, and are frequently adapted to a range of domain-specific applications.

Focus: This Report focuses on ‘emerging risks’: risks that arise at the frontier of AI capabilities. The Bletchley Declaration, issued following the 2023 AI Safety Summit, emphasised that “particular safety risks arise at the ‘frontier’ of AI”, including risks from misuse, issues of control, and cybersecurity risks. The Declaration also recognised broader AI impacts, including on human rights, fairness, accountability, and privacy. This Report aims to complement assessments that consider these broader concerns, including the UN’s Independent International Scientific Panel on AI.*

Independence: Under the leadership of the Chair, the independent writing team jointly had full discretion over its content. The Report aims to synthesise scientific evidence to support informed policymaking. It does not make specific policy recommendations.

* Note that this focus makes the scope of this Report narrower than that of the 2025 Report, which also addressed issues such as bias, environmental impacts, privacy, and copyright.

Process and contributors

The International AI Safety Report is written by a diverse team with over 30 members, led by the Chair, lead writers, and chapter leads. It undergoes a structured review process. Early drafts are reviewed by external subject-matter experts before a consolidated draft is reviewed by:

An Expert Advisory Panel with representatives nominated by over 30 countries and international organisations, including the OECD, the EU, and the UN

A group of Senior Advisors composed of leading international researchers

Representatives from industry and civil society organisations

The writing team, chapter leads, lead writers, and Chair consider feedback provided by reviewers and incorporate it where appropriate.

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.

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.

Introduction

Leading general-purpose AI systems now pass professional licensing exams in law and medicine, write functional software when given simple prompts, and answer PhD-level science questions as well as subject-matter experts. Just three years ago, when ChatGPT launched, they could not reliably do any of these things. The pace of this transformation has been remarkable, and while the pace of future changes is uncertain, most experts expect that AI will continue to improve.

Almost a billion people now use general-purpose AI systems in their daily lives for work and learning. Companies are investing hundreds of billions of dollars to build the infrastructure to train and deploy them. In many cases, AI is already reshaping how people access information, make decisions, and solve problems, with applications in industries from software development to legal services to scientific research.

But the same capabilities that make these systems useful also create new risks. Systems that write functional code also help create malware. Systems that summarise scientific literature might help malicious actors plan attacks. As AI is deployed in high-stakes settings – from healthcare to critical infrastructure – the impacts of deliberate misuse, failures, and systemic disruptions can be severe.

For policymakers, the rate of change, the breadth of applications, and the emergence of new risks pose important questions. General-purpose AI capabilities evolve quickly, but it takes time to collect and assess evidence about their societal effects. This creates what this Report calls the ‘evidence dilemma’. By acting too early, policymakers risk implementing ineffective or even harmful interventions. But waiting for conclusive evidence can leave societies vulnerable to potential risks.

The role of this report

This Report aims to help policymakers navigate that dilemma. It provides an up-to-date, internationally shared scientific assessment of general-purpose AI capabilities and risks.

The writing team included over 100 independent experts, including an Expert Advisory Panel comprising nominees from more than 30 countries and intergovernmental organisations including the EU, OECD, and UN. The Report also incorporates feedback from reviewers across academia, industry, government, and civil society. While contributors differ on some points, they share the belief that constructive and transparent scientific discourse on AI is necessary for people around the world to realise the technology’s benefits and mitigate its risks.

Because the evidence dilemma is most acute where scientific understanding is thinnest, this Report focuses on ‘emerging risks’: risks that arise at the frontier of general-purpose AI capabilities. Its analysis focuses on issues that remain particularly uncertain, aiming to complement efforts that consider the broader social impacts of AI. While this Report draws on international expertise and aims to be globally relevant, readers should note that variation in AI adoption rates, infrastructure, and institutional contexts mean that risks may manifest differently across countries and regions.

The evidence base for these risks is uneven. Some risks, such as harms from AI-generated media or cybersecurity vulnerabilities, now have robust empirical evidence. Evidence for others – particularly risks that may arise from future developments in AI capabilities – relies on modelling exercises, laboratory studies under controlled conditions, and theoretical analysis. The analysis here draws on a broad range of scientific, technical, and socioeconomic evidence published before December 2025. Where high uncertainty remains, it identifies evidence gaps to guide future research.

Changes since the 2025 Report

This edition of the International AI Safety Report follows the publication of the first Report in January 2025. Since then, both general-purpose AI and the research community’s understanding of it have continued to evolve, warranting a revised assessment.

Over the past year, AI developers have continued to train larger and more capable AI models. However, they have also achieved significant capability gains through new techniques that allow systems to use more computing power to generate intermediate steps before giving a final answer. These new ‘reasoning systems’ show particularly improved performance in mathematics, coding, and science. In addition, AI agents – systems that can act in the world with limited human oversight – have become increasingly capable and reliable, though they remain prone to basic errors that limit their usefulness in many contexts.

General-purpose AI systems have also continued to diffuse, faster than many previous technologies in some places, though unevenly across countries and regions. Improved performance in capabilities related to scientific knowledge has also prompted multiple developers to release new models with additional safeguards, as they were unable to confidently rule out the possibility that these models could assist novices with weapon development.

This Report covers all these developments in greater depth, and incorporates several new structural elements to improve its usefulness and accessibility. It includes capability forecasts developed with the Forecasting Research Institute and scenarios developed with the OECD. Each section includes updates since the last Report, key challenges for policymakers, and evidence gaps to guide future research.

How this Report is organised

This Report is organised around three central questions:

What can general-purpose AI do today, and how might its capabilities change? Chapter 1 covers how general-purpose AI is developed (§1.1. What is general-purpose AI?), current capabilities and limitations (§1.2. Current capabilities), and the factors that will shape developments over the coming years (§1.3. Capabilities by 2030).

What emerging risks does general-purpose AI pose? Chapter 2 covers risks from malicious use, including the use of AI systems for criminal activities (§2.1.1. AI-generated content and criminal activity), manipulation (§2.1.2. Influence and manipulation), cyberattacks (§2.1.3. Cyberattacks), and developing biological or chemical weapons (§2.1.4. Biological and chemical risks); risks from malfunctions, including operational failures (§2.2.1. Reliability challenges) and loss of control (§2.2.2. Loss of control); and systemic risks,* including disruptions to labour markets (§2.3.1. Labour market impacts) and threats to human autonomy (§2.3.2. Risks to human autonomy).

What risk management approaches exist, and how effective are they? Chapter 3 covers the distinctive policymaking challenges that general-purpose AI poses (§3.1. Technical and institutional challenges), current risk management practices (§3.2. Risk management practices), the various techniques developers use to make AI models and systems more robust and resistant to misuse (§3.3. Technical safeguards and monitoring), the particular challenges of open-weight models (§3.4. Open-weight models), and efforts to make society more resilient to potential AI shocks and harms (§3.5. Societal resilience).

* 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”.

Many aspects of how general-purpose AI will develop remain deeply uncertain. But decisions made today – by developers, governments, communities, and individuals – will shape its trajectory. This Report aims to ensure that those decisions are made with the best possible understanding of AI capabilities, risks, and options for risk management.

1

Background on general-purpose AI

Over the past year, the capabilities of general-purpose AI models and systems have continued to improve. Leading systems now match or exceed expert-level performance on standardised evaluations across a range of professional and scientific subjects, from undergraduate examinations in law and chemistry to graduate-level science questions. Yet their capabilities are also ‘jagged’: they simultaneously excel on difficult benchmarks and fail at some basic tasks. Current systems still provide false information at times, underperform in languages that are less common in their training data, and struggle with real-world constraints like unfamiliar interfaces and unusual problems. Alleviating these limitations is an area of active research, and researchers and developers are making progress in some areas. Sustained investment in AI research and training is expected to drive continued capability progress through 2030, though substantial uncertainty remains about both what new capabilities will

emerge and whether current shortcomings will be resolved.

This chapter covers current and future capabilities of general-purpose AI. The first section introduces general-purpose AI, explaining how these systems work and what drives their performance (§1.1. What is general-purpose AI?). The second section examines current capabilities and limitations (§1.2. Current capabilities). A recurring theme is the ‘evaluation gap’: how a system performs in pre-deployment evaluations like benchmark testing often seems to overstate its practical utility, because such evaluations do not capture the full complexity of real-world tasks. The final section considers how capabilities might evolve by 2030 (§1.3. Capabilities by 2030). AI developers are investing heavily in computing power, data generation, and research. However, there is substantial uncertainty about how these investments will translate into future capability gains. To illustrate the range of plausible outcomes, the section presents four scenarios developed by the OECD, which range from stagnation to an acceleration in the rate of capability improvements.

1.1. What is general-purpose AI?

Key information

‘General-purpose AI’ refers to AI models and systems that can perform a variety of tasks, rather than being specialised for one specific function or domain. Examples of such tasks include producing text, images, video, and audio, and performing actions on a computer.

General-purpose AI models are based on ‘deep learning’. Modern deep learning involves using large amounts of computational resources to help AI models learn complex relationships and abstract features from very large training datasets.

Developing a leading general-purpose AI system has become very expensive. To train and deploy such systems, developers need extensive data, specialised labour, and large-scale computational resources. Acquiring these resources to develop a leading system from scratch now costs hundreds of millions of US dollars.

Since the publication of the last Report (January 2025), capability improvements have increasingly come from post-training techniques and extra computational resources at the time of use, rather than from increasing model size alone. Previous performance improvements largely resulted from making models larger and using more data and computing power during initial training.

What are general-purpose AI systems?

General-purpose AI systems are software programmes that learn patterns from large amounts of data, enabling them to perform a variety of tasks rather than being specialised for one specific function or domain (see Table 1.1). To create these systems, AI developers carry out a multi-stage process that requires substantial computational resources, large datasets, and specialised expertise (see Table 1.2). Computational resources (often shortened to ‘compute’) are required both to develop and to deploy AI systems, and include specialised computer chips as well as the software and infrastructure needed to run them.* Because they are trained on large, diverse datasets, general-purpose AI systems can carry out many different tasks, such as summarising text, generating images, or writing computer code. This section explains how general-purpose AI systems are made, what ‘reasoning’ models are, and how policy decisions shape general-purpose AI system development.

* The term compute can also refer to either a measurement of the number of calculations a processor can perform (typically measured in floating-point operations per second) or specifically the hardware (such as graphics processing units) that performs those calculations.

Type of general-purpose AI Examples

Language systems

Apertus 1

Claude-4.5 2

Command A 3

EXAONE 4.0 4

Gemini-3 Pro 5

GLM-4.5 6

GPT-5 7

Hunyuan-Large 8

Kimi-K2 9

Mistral 3.1 10

Qwen3 11

DeepSeek-V3.2 12

Image generators

DALL-E 3 13

Gemini 2.5 Flash 14

Midjourney v7 15

Qwen image 16

Video generators

Cosmos 17

SORA 18

Pika 19

Runway 19

Veo 3 20

Robotics and navigation systems

Gemini Robotics 21

Gr00t N1 22

MobileAloha 23 24

Octo 25

OpenVLA 26

PaLM-E

Predictors of diverse classes of biomolecular structures

AlphaFold3 27

AMPLIFY 28

CellFM 29

Evo2 30

AI agents

AlphaEvolve 31

ChatGPT Agent 32

Claude Code 33

Doubao-1.5 34

Magentic-One 35

OpenScholar 36

Table 1.1: There are several different types of general-purpose AI. In this Report, models that can predict structural information for diverse classes of molecules are considered to be ‘general-purpose’ AI because they can be adapted for a variety of tasks. For example, models trained to predict protein structure are applicable to a variety of other tasks, such as predicting protein interactions, predicting small molecular binding sites, and predicting and designing cyclic peptides. 40

Deep learning is foundational to general-purpose AI

Researchers build general-purpose AI models using a process called ‘deep learning’, which trains models to learn from examples. 41 Unlike software engineering, deep learning models learn to accomplish tasks from data instead of relying on hand-written instructions. By processing large amounts of data, such as images, text, or audio, these models discover ways to represent that data, creating internal representations of patterns (such as shapes, word associations, or sound structures) that help the model recognise relationships and generate outputs aligned with its training objective. They then use these learned internal representations as abstract features to analyse new, similar data and generate outputs in the same style. For example, a general-purpose AI model trained on enough examples of 19th-century romantic English poetry can recognise new poems in that style and produce new material in a similar style.

On a more granular level, deep learning works by processing data through layers of interconnected information-processing nodes. These nodes are often called ‘neurons’ because they are loosely inspired by neurons in biological brains (‘neural networks’) (Figure 1.1). 42 As information flows from one layer of neurons to the next, the model progressively transforms the data into more abstract representations as groups of learned features – patterns the model has automatically discovered in the data, rather than hand-coded ones. For example, in an image-processing model, the first layers might learn to detect simple features such as edges or basic shapes, while deeper layers combine these features to pick out more complex patterns such as faces or objects.

The features at all layers are discovered through the optimisation process that defines the training procedure. During training, when the model makes mistakes, deep learning algorithms adjust the strength of various connections between neurons to improve the model’s performance. The strength of each connection between nodes is often called a ‘weight’. This layered approach gives deep learning its name.

Deep learning has proven very effective at allowing AI systems to accomplish tasks that were previously considered difficult for traditional hand-programmed computational systems and other earlier symbolic or rule-based AI methods. Most state-of-the-art general-purpose AI models are now based on a specific neural network architecture known as the ‘transformer’. 43 44 Transformers use an ‘attention’ mechanism 45 that helps the model to focus on the most relevant parts of the input data when processing information, such as determining which words in a sentence are most important for understanding its meaning. This particular way of building models has led to significant improvements in translation, 43 natural language processing, 46 image recognition 47 and speech recognition, 48 49 ultimately leading to the development of today’s most advanced models.

Figure 1.1: An illustrative representation of a ‘neural network’. Today’s general-purpose AI models are based on these networks, which are loosely inspired by biological brains. Different networks have different sizes and architectures. However, all are composed of connected information-processing units called ‘neurons’, where the strengths of connections between neurons are called ‘weights’. Weights are updated through training with large quantities of data. Source: International AI Safety Report 2025 50 (modified).

Figure 1.2: A schematic representation of the stages of general-purpose AI development. Source: International AI Safety Report 2026.

General-purpose AI is developed in stages

Developing a general-purpose AI system involves multiple stages, from initial model training to post-deployment monitoring and updates (Figure 1.2). In practice, these steps often overlap in an iterative manner. Each stage requires different resource inputs (e.g. data, labour, compute) and different techniques, and they are sometimes undertaken by different developers (Figure 1.2 and Table 1.2).

For example, model pre-training generally requires large amounts of compute and data, making this stage particularly sensitive to policies that affect access to computational resources or training data. 51 52 Similarly, data curation and some model fine-tuning methods currently involve large amounts of human labour for initial data

labelling. 53 This stage is therefore sensitive to changes in labour costs, platform policies, or regulations affecting cross-border contracting arrangements.

1. Data collection and curation

Before training a general-purpose AI model, developers and data workers collect, clean, curate, and standardise raw training data into a format the model can learn from. This can be a labour-intensive process. The training datasets behind state-of-the-art models comprise an immense number of examples from across the internet.

Teams often develop sophisticated filtering methods to reduce harmful content, eliminate duplicate data, and improve representation across different topics and sources. 54 55 Data curation can also help reduce copyright and privacy violations, remove examples containing dangerous knowledge, handle multiple languages, and improve documentation for data provenance. 56 57 58

2. Pre-training (first stage of training)

During pre-training, developers feed models massive amounts of diverse data to instil a broad base of information and contextual understanding. This process produces a 'base model'. This is a highly data- and compute-intensive process.

During pre-training, models are exposed to billions or trillions of examples of content such as pictures, texts, or audio. Through this exposure, the model gradually discovers abstract features to represent data and learns about how these features are related, which allows it to make sense of new inputs in context. This pre-training process takes weeks or months 59 and uses tens or hundreds of thousands of graphics processing units (GPUs) or tensor processing units (TPUs) 60 – specialised computer chips designed to rapidly process many such calculations. Some developers conduct pre-training with their own compute, while others use resources provided by specialised compute providers.

3. Post-training and fine-tuning (second stage of training)

'Post-training' further refines the base model to optimise it for a specific application. It is a moderately compute-intensive and highly labour-intensive process. A shift towards using 'synthetic data' – artificially generated information that mimics real-world data but is created using algorithms or simulations – is helping to make this phase less labour-intensive.

Post-training includes various fine-tuning techniques and other modifications. 'Supervised fine-tuning' involves further training a trained model on specific datasets to improve the model's performance in that domain. 61 62 For example, a general-purpose model could be further trained on a large corpus of radiological images. 'Reinforcement learning' (RL) involves improving model performance by 'rewarding' a model (providing positive feedback) for desirable outputs and 'penalising' a model (providing negative feedback) for undesirable outputs. It has two prominent subcategories. 'Reinforcement learning from human feedback' involves rewarding outputs that align with human preferences and penalising those that do not, based on human feedback. 63 64

'Reinforcement learning with verifiable rewards' (RLVR) is used for improving model performance on tasks that require factual correctness, such as maths or code generation. Developers typically alternate between applying post-training techniques and running tests until results show that the model meets desired specifications.

4. System integration

Developers combine one or more general-purpose AI models with other components to create an 'AI system' that is ready for use. GPT-5 (for example) is a general-purpose AI model that processes text, images, and audio, while ChatGPT is a general-purpose AI system that combines several models of different sizes and capabilities with a chat interface, content processing, Web access, and application integration to create a functional product.

In addition to making AI models operational, the additional components in an AI system also aim to enhance capability, usefulness, and safety. For example, a system might come with a filter that detects and blocks model inputs or outputs that contain harmful content. 65 Developers are also increasingly using 'scaffolding' – additional software built around general-purpose AI models that allows them to plan ahead, pursue goals, and interact with the world. 66

5. Deployment and release

Deployment is the process of making the integrated AI system available for its intended use. Developers and deployers implement AI systems into real-world applications, products, or services. Developers can deploy AI

systems internally (for their own use) or externally (for private customers or public use). When deploying AI systems externally, companies often provide users with access through online user interfaces or application programming interfaces (APIs) that allow users to access and run the system. For example, one company might design a bespoke customer service chatbot that is powered by another company’s general-purpose AI system.

‘AI system deployment’ refers to making a model available for real-world use with integrated tools and interfaces, while ‘model release’ involves making the base model accessible to others – either as open-weight (downloadable parameters) or closed-weight (API access only). See §3.4. Open-weight models .

6. Post-deployment monitoring and updates Developers often gather and analyse user feedback, track impact and performance metrics, and make iterative improvements to address issues discovered during real-world use. 67 Improvements are made by updating the system integrations, often via continual fine-tuning and providing models with access to external databases of (recent) facts. This keeps large AI models up-to-date without repeating the full pre-training process. 68 This enables capabilities to accumulate across successive training rounds while maintaining stability and reducing computational costs.

Table 1.2: At each general-purpose AI development stage, the AI model is improved for downstream use and eventually deployed as a fully integrated AI system, monitored and updated.

Reasoning systems generate ‘chains of thought’ during inference to improve performance

Inference happens when someone uses the AI model after it is trained. For example, inference occurs when a person asks an AI system to plan a trip and the model behind it draws on relevant aspects of what it has learned regarding geography, transportation, and cuisine to generate an itinerary.

In the past decade, advances in AI capabilities have largely come from larger training runs; that is, increasing the amount of compute used to train an AI model. Recently, however, researchers have made more gains by allowing models to process information for longer and by training them to produce explicit reasoning steps as they accomplish a task. 69 70 AI systems that work like this are called ‘reasoning systems’, and the intermediate explanations they go through while solving a problem or answering a question are called ‘chains of thought’. Reasoning systems require more computational resources at the time of use to generate these sophisticated chains of thought, 71 72 73 74 and more resources during training so that they learn to reason better. In practice, these reasoning capabilities let AI systems solve more complex problems by iteratively decomposing a task into smaller steps. Table 1.3 shows an example of a non-reasoning system and a reasoning system solving the same problem.

Reasoning systems have achieved major breakthroughs in capabilities on challenging problems. For example, in 2025, reasoning systems specialised for mathematical problem-solving, such as Google’s Gemini Deep Think and an unreleased, experimental model from OpenAI, solved International Mathematical Olympiad problems (in a structured test setting) at a level equivalent to human gold-medal performance. 75 76 Reasoning systems have demonstrated significant progress in formal domains such as mathematics, logic puzzles, and structured scientific questions, where step-by-step reasoning can be explicitly verified. 77 However, reasoning systems can also fail by producing irrelevant, unproductive, or repetitive chains of thought. 78 79

Updates on training methods

Since the publication of the last Report (January 2025), a training method called ‘distillation’ has greatly increased the efficiency with which some models can be fine-tuned. Distillation involves training a ‘student’ model on the outputs of a more powerful (and usually larger) ‘teacher’ model, allowing the student model to directly imitate the outputs of the teacher. 80 For example, DeepSeek developed a large model called DeepSeek-R1, which excels at chain-of-thought reasoning. R1 produced reasoning outputs which were then used to fine-tune smaller student models, including DeepSeek-V3. DeepSeek-V3 maintains much of R1’s mathematical, coding, and document-analysis capabilities and was reportedly fine-tuned for approximately \$10,000 USD (though its pre-training costs were not reported). 81 This is likely orders of magnitude lower than the cost of fine-tuning similarly capable, larger models.

Table 1.3: An example of a non-reasoning system (left) versus a reasoning system (right) solving the same riddle. These examples are adapted from real AI responses. The reasoning system spends more time and computational power on ‘thinking’ by constructing a ‘chain of thought’ before providing its final answer.

Figure 1.3: An illustrative representation of an AI agent: an AI model (centre) that has been configured to

iteratively plan, reason, and use tools to accomplish real-world tasks. Source: International AI Safety Report 2026

Distillation can thus be a cheap and efficient way for models to gain more powerful capabilities. 82 Some researchers have used distillation to fine-tune highly capable models using as few as 1,000 examples generated from state-of-the-art models. 83 Since distillation requires a pre-existing teacher model, it cannot be directly used to advance state-of-the-art model capabilities. However, it can speed up the proliferation of advanced AI capabilities, even from closed-source models. 84

Together with technological advances in ‘distributed compute’ and decentralised training (approaches where developers use multiple processors, servers, or data centres working together to perform AI training or inference, 85 86 87 the degree to which many AI development projects depend on large-scale, centralised compute infrastructure has been reduced. This increasingly enables less well-resourced actors to develop and deploy powerful systems.

Updates on AI agents

Since the last Report (January 2025), advances in how developers combine AI models with tools have enabled the development of increasingly powerful AI agents. AI agents are designed to pursue goals, which are often specified by users in natural language. To achieve these goals, they are given access to tools, such as memory, a computer interface, and web browsers. These tools and the code used to combine them with the model are referred to as ‘scaffolding’, and they help AI agents autonomously interact with the world, make plans, remember important details, and pursue goals 88 89 with much less oversight or assistance from humans. For example, Manus AI is a popular AI agent that can automate various tasks, including Web search, software development, and online purchases. 90 Figure 1.3 illustrates a simple example of an AI agent composed of a general-purpose AI model ‘brain’ that can iteratively plan, reason, and use tools for memory, web browsing, and computer use.

Digital infrastructure for AI agents is expanding, 91 and they are increasingly common across industries. 92 93 94 AI agents have been developed for tasks such as research, 37 software engineering, 95 robotic control, 96 and customer service. 97 Ongoing research and development has resulted in steadily more capable and more autonomous AI agents or multi-agent systems. Researchers have estimated that the complexity of software benchmark tasks that AI agents can accomplish doubles approximately every seven months (see also §1.2. Current capabilities). 98 Experts argue that increasingly capable AI agents will give rise to both major opportunities and risks 99 100 (see §2.2.1. Reliability challenges).

Evidence gaps

The main evidence gaps around the general-purpose AI system development process stem from a lack of publicly available information regarding how they are developed. Some developers are highly transparent about how they develop general-purpose AI systems. 1 101 However, in general, there is a limited degree of public and policymaker knowledge about how most advanced models are developed, safeguarded, evaluated, and deployed. This is particularly true for internally deployed AI systems that are used within AI companies but not used or understood by outside stakeholders. 102 103 This limited external visibility creates challenges for transparency and oversight. Various researchers have pointed to limited and inconsistent transparency around training data, 104 105 106 general-purpose AI models, 107 108 AI agents, 92 evaluations, 109 development pipelines, 110 and safety. 111 Limitations to external disclosure are sometimes necessary to protect companies’ trade secrets and intellectual property. At the same time, low transparency makes it more difficult for independent researchers and policymakers to study general-purpose AI models and systems.

Close

1.2. Current capabilities

Key information

General-purpose AI systems can perform a wide range of well-scoped tasks with high proficiency. These include conversing fluently in numerous languages; generating code to complete narrow software tasks; creating realistic images and short videos; and solving graduate-level mathematics and science problems.

However, their capabilities are ‘jagged’: there remain many tasks AI systems do not perform well. For example, AI systems can be derailed by simple errors during multi-step projects; continue to generate text that includes false statements (‘hallucinations’); and cannot yet integrate with robotic components to perform basic physical tasks

such as housework. Their performance also tends to decline when prompted in languages other than English, which are less represented in training datasets.

AI agents are increasingly able to do useful work. For example, AI agents have demonstrated the ability to complete a variety of software engineering tasks with limited human oversight. However, they cannot yet complete the range of complex tasks and long-term planning required to fully automate many jobs.

Since the publication of the last Report (January 2025), advances in ‘reasoning systems’ have driven performance improvements on more complex tasks. Reasoning systems are able to break problems into smaller steps and compare alternative answers. This has especially improved their performance on tasks related to mathematics, coding, and scientific research.

A central challenge is an emerging ‘evaluation gap’: existing evaluation methods do not reliably reflect how systems perform in real-world settings. Many common capability evaluations are outdated, affected by data contamination (when AI models are trained on the same questions used in evaluations), or focus on a narrow set of tasks. As a result, they provide limited insight into real-world AI performance.

General-purpose AI systems exhibit many remarkable capabilities. Leading systems now perform at gold-medal level in mathematics competitions and assist scientific researchers with generating hypotheses and troubleshooting laboratory work. They match, and in some cases exceed, expert performance on a wide range of benchmarks and task-specific evaluations. Yet the performance profile these systems display is also ‘jagged’: their capabilities vary widely among different tasks and contexts. They still sometimes generate false information (‘hallucinations’) and produce inconsistent outputs even when given identical or similar inputs. An ‘evaluation gap’ exists: AI systems often perform impressively in controlled settings such as pre-deployment evaluations, but more poorly in real-world conditions. This variability makes it difficult to assess general-purpose AI capabilities with a single metric. This section outlines both the capabilities of AI systems and their shortcomings (Table 1.4).

Table 1.4: A summary of the main capabilities and limitations of current general-purpose AI systems.

What can current general-purpose AI systems do?

General-purpose AI systems now perform at or above the level of human experts on standardised evaluations, covering a growing range of well-defined professional and scientific subjects (Figure 1.4). For example, leading models score over 90% on undergraduate-level examinations in subjects from chemistry to law (MMLU, 112) and achieve over 80% on graduate-level science tests (GPQA, 14). In July 2025, models from Google DeepMind and OpenAI reached gold medal-level scores at the International Mathematical Olympiad, solving five out of six problems under competition-like conditions. 76 Beyond text-based reasoning, these systems display powerful multimodal capabilities: they can create photorealistic images, short high-definition videos, 3D scenes, and musical pieces from simple text prompts, 13 18 113 114 115 116 and they are beginning to process complex sensor data to guide physical robots. 21

Advanced capabilities are increasing productivity in medicine, education, software development, and other sectors

Advanced AI capabilities now power practical tools that match or exceed human performance on specific tasks, increasing productivity in multiple sectors. 117

In medicine , AI systems can analyse clinical scenarios and conduct diagnostic conversations to generate lists of potential diagnoses. In specific simulated settings, their accuracy can exceed that of human physicians, 118 119 though they lack the reliability and consistency required for real-world clinical deployment.

In education , AI systems are being rapidly adopted in areas from curriculum design to student assessment, transforming the education process, 120 121 while widespread use by students is posing significant challenges to the integrity and validity of existing academic assessments. 122

In software development , AI coding assistants are now widely adopted, with some studies suggesting that developers using AI assistants complete certain tasks 20–30% faster on average than those without. 123 124 125

Large-scale studies in other sectors such as customer service, consulting, and professional writing find measurable productivity gains from AI-assisted work, though these effects vary across tasks and worker groups. 126 127 128 129 130 (For a more detailed discussion of the labour market implications of general-purpose AI, see §2.3.1. Labour market impacts .)

General-purpose AI systems assist scientific research

General-purpose AI systems are now used by researchers to support relatively complex tasks across disciplines. Researchers have demonstrated that AI systems can, under high-level human guidance, design novel proteins for medical use, which are later validated in a physical laboratory. 131 Other systems have discovered new algorithms that are more efficient than long-standing human-designed methods. 31 Notably, such advances often rely less on the raw power of the latest models and more on appropriate system integration. General-purpose AI is also increasingly used to accelerate AI research itself, a trend with significant implications discussed further in §1.3. Capabilities by 2030 . In the social sciences, researchers are using AI to accelerate data analysis through automated annotation and to explore social dynamics by simulating individual and collective behaviour with AI agents. 132 133 134 Moving from analysis to direct application, researchers are beginning to use general-purpose AI systems to design and study scalable, novel social interventions. For example, recent work has explored using AI-mediated conversations to find common ground in democratic debates or to reduce belief in conspiracy theories through dialogue. 135 136 137

Figure 1.4: 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. 138

What are the current limitations of general-purpose AI systems?

Despite advances in capabilities, the performance of general-purpose AI systems remains jagged across tasks and contexts. This section highlights some prominent limitations, though the full range of challenges is broader.

Reliability challenges persist in current AI systems

Despite recent improvements, general-purpose AI systems can be unreliable and prone to basic errors of fact and logic. Even systems that excel at complex tasks may generate non-existent citations, biographies, or facts – a phenomenon known as ‘hallucination’. 139 140 141 Their performance can also be inconsistent; for example, accuracy on maths problems can decrease significantly when irrelevant information is inserted into the problem description. 142 This brittleness extends to multimodal capabilities, where models often have low performance on spatial reasoning tasks, such as basic counting of objects in a scene. 143 144 While expert human oversight can mitigate some of these risks, there is a corresponding danger of over-reliance, where users trust incorrect outputs because they are presented fluently and confidently 145 146 (see §2.3.2. Risks to human autonomy). This unreliability makes it difficult to safely adopt such systems in high-stakes settings such as medicine and finance, where errors can have grave consequences, and human verification of system outputs remains necessary.

Systems struggle with long-term planning and unexpected obstacles

General-purpose AI systems also struggle with tasks that require long-term planning, maintaining a coherent strategy over many steps, and adapting to unexpected obstacles. As tasks grow longer, AI agents often lose track of their progress and cannot reliably deal with unexpected inputs. 147 148 149 For example, even a simple website pop-up ad can derail an entire task. 150 Large-scale evaluations confirm this pattern: in software development, the most capable systems achieve only 50% success on tasks lasting just over two hours, and reaching 80% success requires limiting them to much simpler 25-minute tasks. 98 151 For now, reliable automation of long or complex tasks remains infeasible.

Interacting with the physical world remains challenging

Progress on digital tasks has also proved difficult to translate into robotics, where the complexity of the physical world introduces new challenges. Recent advances are centred on Vision-Language-Action (VLA) models – foundation models designed to enable robots to follow natural language instruction, interpret multimodal sensory data, and generate motor commands. State-of-the-art systems such as π 0.5 152 and Gemini Robotics 21 can now interpret simple verbal commands such as ‘clean the kitchen’ and execute a sequence of physical steps in an unfamiliar, controlled environment. However, current VLA models still do not perform well with unusual object shapes and unexpected events. 152 Ensuring that such systems can operate safely and reliably to minimise the risk of physical harm or property damage, and perform well in diverse environments remains an active area of research. 153 154 155

Performance is uneven across languages and cultures

The capabilities of general-purpose AI models and systems also vary across languages and cultures. Performance is highest on tasks in English, reflecting the fact that most training data comes from Western sources. 156 157 For example, one evaluation of AI models across 83 languages found substantially lower performance on languages that use non-Latin scripts and on languages with limited digital resources. 158 This disparity extends to cultural knowledge; 159 in one study, AI models correctly answered 79% of questions about everyday US culture but only 12% of questions about Ethiopian culture. 160 Another study finds that current models ‘reason’ more effectively in high-resource languages, which may widen the performance gap between languages. 161 Beyond language and culture, similar patterns appear along geographic and socioeconomic lines. Models underrepresent locations with disadvantaged demographics in recommendations 162 – for example, if asked for a restaurant recommendation, they might fail to suggest restaurants in poorer areas – and their performance on factual recall degrades for lower-income countries. 163 164 This inequality is compounded by evaluation benchmarks that are themselves heavily skewed toward English, creating an ecosystem where low-resource languages remain systematically understudied and underoptimised. 165 166

Updates

Since the publication of the last Report (January 2025), ‘reasoning’ systems have become mainstream (see §1.1. What is general-purpose AI? for details of their development). These systems demonstrate substantially improved performance on hard mathematics, coding, and scientific tasks by generating and comparing multiple answers within their own ‘chain of thought’ before producing a final answer (Figure 1.5). 112 167 Because these models’ performance depends in part on inference-time compute, their effective capabilities can change after initial development – improving as more computational resources are allocated.

In parallel, AI companies have focused more on developing AI agents, especially in areas such as software engineering 168 and computer use. 169 170 While reliability remains a bottleneck, the complexity of tasks these agents can automate has been increasing rapidly. 98 Finally, enabling models to form long-term memories and learn continuously from user interaction is emerging as a key area of development. 171 172

Figure 1.5: Performance of a general-purpose AI model (s1) on reasoning-intensive tasks with varying amounts of test-time compute (i.e. when using additional compute during inference). Allocating more computational time during response generation leads to substantially better results on mathematics (AIME 24) and PhD-level science questions (GPQA Diamond). Source: Muennighoff et al., 2025. 173

Evidence gaps

Jagged capabilities and the evaluation gap make general-purpose AI capabilities difficult to reliably measure and predict. 174 175 Performance also depends heavily on the specific test examples and prompt used, making it difficult to prove with high-confidence that an AI system cannot perform certain – potentially dangerous – tasks. 176 There is no single, comprehensive, and continuously updated synthesis of AI capabilities, leading to a fragmented and often outdated understanding of the field. Existing reviews, 138 177 including this Report, provide valuable summaries but are static snapshots in a rapidly moving field. With no widely accepted taxonomy for capabilities, policymakers must navigate a patchwork of benchmarks and sources to form a complete picture.

Benchmarks often fail to predict real-world performance

Benchmark integrity is a growing concern. Many capability evaluations rely on standardised benchmarks. However, many models may have been trained using data from these same benchmarks – a problem called ‘data contamination’, which most developers do not currently track or disclose. 178 This can lead to inflated performance scores that do not reflect a model’s true ability, 179 but rather its capacity to memorise answers. 180 181 182 A further limitation of current evaluation practices is that they rely on automated testing in controlled lab environments. However, this often overestimates AI systems’ practical utility in dynamic, real-world settings. 147 149 183 184 For example, one study found that, while an AI agent could produce functional code, the code still required significant human effort to fix issues with documentation, formatting, and quality before it was usable in a real project. 185 To address these limitations, a dedicated ‘evaluation science’ is emerging, advocating for rigorous methodologies that ensure external validity and better predict real-world performance. 186 187 For instance, recent benchmarks have begun to measure AI system performance on economically valuable tasks 188 189 and real-world remote labour. 190 191

The evidence for how AI augments human capabilities is inconclusive

Measuring AI's practical benefits consistently is challenging because success depends on both the specific task and the user's skill at leveraging AI for it, meaning lab results often fail to predict real-world value. For example, one study shows that a model's standalone accuracy is not a reliable predictor of human-AI team performance.¹⁹² Many studies confirm positive uplift from using AI.^{126 127 128} However, one recent study found that, although software developers believed that AI was making them more productive, it actually slowed down experienced programmers by 19% on complex coding tasks.¹²⁹

Close

1.3. Capabilities by 2030

Key information

Investments in AI development are expected to grow significantly in coming years. Forecasts suggest that the computational power used to train the largest AI models could grow 125-fold by 2030 without hitting hard limits in energy, chips, or data. Training methods are also projected to use that computing power two to six times more efficiently each year.

Plausible trajectories for capability improvements range from incremental or even plateauing progress to rapid acceleration. Uncertain factors such as technical limits or energy bottlenecks could constrain capability gains despite large investments, while positive feedback loops – such as AI systems contributing to AI research – could accelerate progress. There is little expert consensus on which trajectory is most likely.

If capabilities continue to improve at their current rate, by 2030 AI systems will be able to complete well-scoped software engineering tasks that would take human engineers multiple days to complete. Projections for future performance in other domains are scarce, and the extent to which capability improvements will generalise to domains where training data is more limited and performance hard to assess is unclear.

Since the publication of the last Report (January 2025), key trends suggest that capabilities will continue improving. In expectation of future gains, AI companies have announced unprecedented investments of more than \$100 billion in data centre development to support larger training runs and wider deployment.

Beyond 2030, the trajectory of AI capabilities becomes even harder to forecast. Over time, some experts expect it will be harder to obtain data, chips, capital, and energy at the scale needed for larger training runs. However, researchers may find ways to use these resources more efficiently or discover new approaches that sidestep current bottlenecks. Which considerations will prove most important is highly uncertain.

The key inputs of AI progress – compute, algorithmic improvements, and data – have grown exponentially in recent years, and new inference-time scaling methods are further improving models' capabilities, even after they are trained. If these trends continue, experts expect AI capabilities to advance substantially by 2030. However, researchers cannot reliably predict when specific capabilities will emerge, and experts disagree about whether exponential increases in inputs will continue. Some expect that current training techniques will plateau, or that bottlenecks in data and energy will limit future progress. Yet others think that progress will accelerate further, since the application of AI systems to AI research itself could create positive feedback loops.^{193 194} To illustrate these divergent trajectories, this section presents four AI capability scenarios for 2030, developed in collaboration with the OECD. Additional technical details on scaling laws, inputs to scaling, and current benchmark performance are provided in the technical supplement.

Drivers of progress: compute, algorithms, and data

Frontier AI progress is driven by three inputs: compute, algorithmic advances, and data.

Compute refers to the computational resources, including hardware, software, and infrastructure, used in AI development and deployment. More compute allows for larger models to be trained on larger datasets (Figure 1.6), leading to better performance across various tasks.^{195 196} Compute can also be used during deployment to improve the quality of an AI system's outputs.^{197 198}

Algorithmic advances improve how efficiently computational resources translate into model performance, and they can also enable qualitatively new capabilities. One model is more efficient than another if it uses less training or inference compute to reach the same performance.¹⁹⁹ For example, GPT-5 is more efficient than GPT-4.5, because it was likely trained with less compute,²⁰⁰ but it outperforms 4.5 on a range of benchmarks, such as GPQA Diamond, which features PhD-level science questions.²⁰¹

Data refers to the information used to train models, including text from the internet, images, and artificially generated synthetic data. 202 Both the amount and the quality of data are relevant for progress.

Figure 1.6: The amount of compute, measured in floating point operations (FLOP), used to train leading AI models between 2012 and 2025. The largest training runs have now likely exceeded 10²⁶ FLOP. Source: Epoch AI, 2025. 203

Figure 1.7: 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. 98

In recent years, all three drivers have grown dramatically. For the most compute-intensive models, training compute has grown about 5x per year. If this trend were to continue until 2030, these models could be trained with roughly 3,000 times more compute than they are today. 204 205 Algorithmic efficiency, according to a 2024 study, has improved roughly 2-6x per year, reducing the compute needed for equivalent performance. 199

Training datasets have expanded from billions to trillions of data points, with an average 2.5x annual increase. 206 New inference-time scaling methods further improve capabilities once a model is trained, unlike traditional approaches that depend mostly on more training compute and larger datasets. 173 207 One study finds that AI systems can complete well-specified software engineering tasks that take human experts 30 minutes around 80% of the time, and the duration of these tasks has been doubling every seven months (Figure 1.7). If this trend continues, AI systems could complete tasks lasting several hours by 2027 and tasks lasting several days by 2030. 98

How will AI capabilities change in the coming years?

Exponential growth in key inputs until 2030 is technically feasible

Exponential growth in key inputs to frontier AI – compute, algorithmic techniques, and data – is technically feasible until around 2030. Analyses of constraints such as production capabilities, investment, and technological progress suggest that compute per frontier model could continue growing at current rates without hitting fundamental bottlenecks in chip manufacturing or energy production. 204 208 To support this scaling, companies are making large investments in compute infrastructure; for example, Meta and OpenAI have announced plans to spend \$65 billion and \$500 billion respectively. 209 210 Importantly, these investments also support increases in inference compute and computational resources for research and development (R&D), the latter of which constitutes the bulk of AI company compute spending. 211

Algorithmic efficiency improvements have historically provided an additional 2–6x performance gain per year. 199 However, experts disagree about how sustainable this growth is, especially beyond 2030. Disagreement centres on whether energy constraints and high-quality data scarcity will force fundamental changes to current development approaches. 206

Experts expect progress in problem-solving to continue

As discussed in §1.2. Current capabilities, AI models have made rapid advances in mathematical reasoning. Building on these advances, experts forecast major progress in reasoning-based problem-solving by 2027–2028. In a study by the Forecasting Research Institute, experts forecast a 50% chance that AI models will achieve 55% accuracy by 2027 and 75% accuracy by 2030 on undergraduate-level FrontierMath problems. 212 However, experts disagree on whether these capabilities will generalise beyond mathematics and programming. Most evidence on the impact of reasoning techniques remains restricted to these domains. 197 213 214 More extensive evaluations and attempts at applying AI systems' reasoning skills to novel domains, such as legal and scientific reasoning, will be required to determine how far reasoning techniques will generalise.

AI systems have also made rapid gains in autonomous software execution. AI systems that could only complete tasks taking human experts a few seconds in 2019 can now, with an 80% success rate, finish software engineering tasks that take human experts 30 minutes. 98 215 This metric – the maximum task duration that AI systems can complete with an 80% success rate – has been doubling roughly every seven months for the past six years. If it were to continue, AI systems could autonomously complete hours-long software projects by 2027–2028 and days-long projects by the end of the decade. However, these projections assume an 80% success rate, which likely falls below the standards required for autonomous deployment in many professional settings. Current evidence shows declining performance as tasks get longer, suggesting that achieving a production-ready success rate may require new innovations. 98 Additionally, the benchmark tasks differ systematically from real-world

software work in ways that may overstate progress: for example, they do not feature ‘messy’ real-world features such as resource constraints, incomplete information, or multi-agent coordination. 98

Experts disagree on the scale and timing of advances in specialised domains

General-purpose AI capabilities are expected to improve across many specialised domains by 2028–2030, though experts disagree about the extent and timing of these advances. AI systems have already surpassed graduate-level performance on some scientific benchmarks, such as GPQA Diamond, where leading models now exceed PhD-level experts. 216 Trend extrapolations suggest that models could reach research-level performance across specialised scientific domains in the next few years, although forecasts remain uncertain.

Specific capabilities can emerge unpredictably even as overall performance improves steadily. For example, general-purpose AI models showed a sharp performance jump in adding large numbers once they were prompted to work step-by-step, rather than gradually improving at this as models scaled. 217 218 219 220 221

Researchers refer to such sudden jumps as ‘emergent capabilities’. These create planning challenges because it is difficult to anticipate when AI systems will suddenly acquire strategically relevant cognitive abilities. Importantly, researchers cannot yet determine whether new prediction methods will make capability emergence more forecastable, and they disagree on how unpredictable these capability leaps truly are. 222 223 224 225

What bottlenecks might slow down progress?

Economic returns from additional compute may diminish

Resource scaling alone may lead to diminishing economic returns and threaten to slow progress, since ever-larger investments will be required to sustain the same rate of capability improvements. Current frontier AI training runs already cost approximately \$500 million in computational resources alone, with next-generation models projected to require \$1–10 billion. 204 226 Meanwhile, consumer trust in AI systems is still low on average, and many enterprises are struggling to adopt AI systems successfully, making large-scale investments of hundreds of billions of dollars a bet on uncertain returns. 93 209 227 If such investments fail to generate revenue (Figure 1.8), companies may sharply reduce scaling investments. This would create a potential ceiling on capability progress, since without continued investments, the 5x annual increase in training compute that has been a driver of recent advances would slow substantially. In that case, capability gains would depend more heavily on algorithmic progress rather than physical scaling alone.

Figure 1.8: Estimated annualised revenue of major AI companies since 2023. Source: Epoch AI, 2025. 228

It is unclear how much AI-assisted research automation will accelerate AI R&D

Experts disagree about whether AI-assisted research automation could dramatically accelerate AI progress in the coming decade. In a pilot study, forecasting experts were asked about the probability that progress in the next few years could compress six years of advancement (2018–2024) into just two years. AI forecasting experts gave a median 20% probability, while superforecasters (skilled generalist forecasters) estimated only 8%. However, forecasters’ estimates increased to 18% in scenarios where AI systems perform better than human researchers on month-long research projects. 229 In such scenarios, AI research could become fully automated much sooner, which some have hypothesised could greatly accelerate AI progress.

Current empirical evidence on AI-assisted research automation is mixed. On a benchmark measuring AI research engineering capabilities, AI agents perform better than humans at two-hour tasks but have lower success rates at eight-hour tasks. 230 While suggestive, this evidence does not account for real-world bottlenecks in AI R&D, such as the fact that researchers must manage ambiguous goals, and that it takes a long time to learn whether an algorithmic improvement actually improved model performance. This uncertainty creates extreme planning challenges for policymakers and institutions: if each AI advancement that accelerates the pace of AI R&D also facilitates the next advancement, decades of progress could happen in years.

Commercial deployment often lags behind capability improvements

Current AI systems demonstrate advanced capabilities in controlled settings, but their adoption occurs at different speeds across sectors. AI coding assistants achieved widespread adoption among software developers within one to four years of release. 231 In contrast, many sectors face substantial obstacles to deploying AI systems. 232 233 Healthcare AI systems that achieve human-level diagnostic accuracy in research settings often require three to five additional years for regulatory approval, clinical integration, and physician training before widespread deployment. 234 Experts forecast that deployment of autonomous vehicle technology will still be limited in 2030,

citing barriers including cultural resistance, infrastructure requirements, and regulatory pushback. 212 Small and medium enterprises, which employ 60% of workers globally, face particular deployment challenges including limited technical expertise, insufficient computational infrastructure, and prohibitive integration costs that can delay AI adoption. 235 236 Geopolitical factors, including export controls on advanced semiconductors and divergent regulatory frameworks across jurisdictions, could create additional barriers that affect both the development and deployment of AI capabilities. 237 238

That said, experts disagree about whether deployment gaps will narrow quickly or persist as a long-term constraint. On the one hand, the rapid uptake of AI tools across particular sectors suggests that deployment will accelerate if organisations observe concrete productivity gains and competitive advantages. 239 Other researchers contend that organisational and regulatory adaptation inherently takes years, regardless of technical progress. 240 This disagreement has implications for policy timing. Policies designed for rapidly deployed AI capabilities may be premature, while those assuming slow adoption may be insufficient to properly manage the risks.

Close

What could progress through 2030 look like: OECD progress scenarios

Considering current trends and uncertainties, including those detailed above, the OECD has developed expert- and evidence-informed scenarios for how AI could advance – or slow down – by 2030. 241 The OECD collaborated with the International AI Safety Report to integrate these scenarios into the Report. The analysis suggests that four broad classes of scenarios are all plausible by 2030:

Scenario 1: Progress stalls

A scenario in which AI capabilities remain largely unchanged. Rapid gains observed over recent years halt, and progress plateaus.

Scenario: In 2030, AI systems can quickly undertake a range of tasks that would take humans hours to perform, but issues of robustness and hallucinations impact reliability. 98 242 AI systems typically rely upon substantial support from humans to complete tasks, such as detailed prompting, review, and provision of context. They lack robust abilities to learn new skills or form memories, maintain coherence over longer complex tasks, or engage with dynamic physical or social environments. 243

Pathway: After 2025, gains within existing approaches for developing frontier AI models hit fundamental limits. This could occur if AI progress slows due to: diminishing returns from larger training runs and more powerful reasoning systems; limitations in accessing computing resources or other critical inputs; a significant drop in AI investment; or the absence of major algorithmic breakthroughs. 244 245

Historical analogue: Passenger aircraft speed, which climbed quickly from 1930 to 1960 before levelling off at 500 knots due to practical limitations. 246

Scenario 2: Progress slows

A scenario in which incremental gains within existing approaches to training AI systems deliver continued but slower progress.

Scenario: In 2030, AI systems are comparable to useful assistants. They have a deep knowledge base, excel at standard forms of structured reasoning, and can usefully perform tasks that require them to use a computer, navigate the Web, or undertake limited interaction with people or services on behalf of the user. They can retain relevant memories, maintain coherent thinking, and error-correct to perform longer or more complex tasks. They lack robust abilities to learn new skills and can handle physical or embodied social tasks only in limited, controlled environments (such as factories or laboratories).

Pathway: After 2025, the approaches of frontier model developers struggle to overcome limitations in continual learning, metacognition and agency, problem-solving, creativity, physical tasks, and social interaction, with existing training paradigms providing imperfect solutions. 243 Scaling of pre-training, inference and post-training combined with some algorithmic innovations continue to deliver progress, but it is slower than in recent years and reasoning systems fail to generalise as well as hoped. 247 248 The ability to continue scaling is slowed as investors see lower returns from continued investments. Bottlenecks in hardware, infrastructure, natural resources, data supply, and energy limit the ability to rapidly scale compute and data. 208

Historical analogue: Antibiotic discovery, which saw a ‘golden era’ of rapid breakthroughs from the 1940s to 1960s, then slowed as the low-hanging fruit from existing discovery methods were exhausted. 249

Scenario 3: Progress continues

A scenario in which continued rapid progress occurs.

Scenario: In 2030, AI systems are comparable to expert collaborators. They can successfully perform many professional tasks in digital environments that might take humans a month to complete. AI systems typically rely upon humans to provide high-level directions, but can often work with high autonomy towards a given objective, including autonomously interacting with a range of stakeholders. They can effectively form and retrieve memories and can ‘learn on the job’ to some extent. They can successfully handle some physical tasks and embodied social tasks beyond controlled environments.

Pathway: After 2025, AI capabilities continue to grow rapidly through larger training runs, more powerful reasoning systems, and new algorithmic innovations. 151 Compute and data inputs continue to scale and do not hit substantial limits before 2030, matching current estimates of the possible scope for continued growth. 203 208 Iteration and extension of existing approaches or novel algorithmic innovations enable developers to overcome current limitations in areas such as continual learning.

Historical analogue: Moore’s law, where computing power on chips doubled approximately every two years over five decades. 250

Scenario 4: Progress accelerates

A scenario in which dramatic progress leads to AI systems as or more capable than humans across most or all capability dimensions.

Scenario: In 2030, AI systems are comparable to human-level remote workers. AI systems’ autonomy and cognitive capability match or surpass humans in cognitive tasks. They capably and autonomously work towards broad strategic goals that they can reflect upon and revise if circumstances change, while also collaborating with humans where necessary. AI systems can seamlessly learn new information and skills during deployment. AI-guided robots can handle complex physical or social tasks in dynamic real-world environments in many industries and roles. AI performance still largely lags humans’ in these physical and embodied tasks, unless the system was developed specifically for a given task, due to challenges in generalisation across physical tasks. 251 252

Pathway: After 2025, there are continued exponential gains in AI capabilities within existing paradigms via continued or accelerated scaling of pre-training, post-training, and inference. These are amplified by significant algorithmic breakthroughs and increasingly substantial contributions from AI coding assistants to the development of AI. 31 253

Historical analogue: DNA sequencing saw superexponential improvements from 2000 to 2020 due to the development of new sequencing paradigms. 254

This scenario analysis suggests that, by 2030, AI progress could plausibly range from stagnation to rapid improvement to levels that exceed human cognitive performance. The full analysis supporting these scenarios is available in OECD (2026) AI Progress Scenarios: Exploring possible pathways through 2030 . 241

Close

Updates

Since the publication of the last Report (January 2025), observed developments have largely remained consistent with the rapid AI progress trajectories outlined in that Report. General-purpose AI systems have become substantially more capable, affordable, and widely adopted, with particularly notable advances in scientific reasoning and autonomous task execution. Major AI companies and cloud providers have announced unprecedented data centre investments totalling hundreds of billions of dollars, demonstrating sustained commitment to the compute scaling trends anticipated in the previous Report. 255 256 257 AI developers have made substantial progress in developing agents that can more reliably execute longer multi-step tasks with reduced human oversight, including advancements in computer use and tool use. The adoption of inference-time compute scaling has become widespread across multiple developers. 167 258 259 260 AI tools are now routinely integrated into AI development workflows for writing training code, designing hardware architectures, and generating synthetic training data.

Evidence gaps

The main evidence gaps around future AI capabilities include limited scientific evidence relevant to forecasting, insufficient data about real-world constraints on AI progress, and limited understanding of whether and to what extent automation could accelerate AI development. First, researchers cannot reliably predict when AI systems will have certain capabilities, or where diminishing returns to scaling key inputs will constrain progress. The relationship between benchmark performance and real-world performance also remains poorly understood; so even if benchmark performance was easily predictable, the associated real-world impacts would be highly uncertain.

Second, there is limited evidence around the real-world constraints that could limit AI progress. These constraints include unclear availability of training data beyond 2030 and whether energy production, chip manufacturing, and capital expenditures can keep pace with the demands of AI development.

Third, there is minimal empirical understanding of feedback loops from AI automating its own research and development. 194 In particular, there are major uncertainties about how much human oversight will be needed in this process, and about whether slow feedback loops in large-scale experiments could constrain acceleration. 261 262

These evidence gaps force policymakers to navigate between two pitfalls: underestimating rapidly emerging capabilities on the one hand, and overreacting to technical advances that may not translate into practical applications on the other. This makes contingency planning across multiple scenarios essential.

Challenges for policymakers

For policymakers working on AI capability forecasting, key challenges include unreliable measurement tools and uncertainty about when certain capabilities will be developed. Current benchmarks often fail to accurately represent real-world capabilities, prompting increased efforts to develop more challenging and realistic evaluations. 263 264 265 266 For example, even if a model achieves 90% accuracy on a programming benchmark, this does not imply that it can build functional software applications. Estimates of algorithmic efficiency progress are highly uncertain due to limited data on key indicators, such as training efficiency improvements, inference-time optimisations, and architectural innovations. For example, while studies of algorithmic efficiency in language models suggest efficiency improvements of 3x per year based on previous data points, they are unable to rule out rates ranging from 2–6x per year. 199

This forecasting problem compounds the uncertainty about capability trajectories, which have vastly different policy implications. If algorithmic progress continues at the upper bound of current estimates, models could achieve equivalent capabilities with 10–100x less compute by 2030. Regulators will therefore need to consider frameworks that can adapt or remain robust to rapid changes in the rate of AI progress and in what AI development looks like, particularly in terms of the required resources. To reduce uncertainty, it will be important to monitor concrete indicators including real-world task evaluations, the rate of algorithmic innovation, and the emergence of qualitatively new capabilities.

Close

Technical supplement

Scaling laws are often used as empirical guidance

‘Scaling laws’ describe predictable relationships between model size, computational resources, and performance. When model developers increase training compute by 10x, model performance tends to improve by a predictable amount across diverse tasks such as language understanding, image recognition, and code generation. 195 196 This predictable relationship has held across six orders of magnitude of model size – from small research models to today’s frontier AI systems, which cost hundreds of millions of dollars to train – suggesting that these patterns reflect fundamental properties of how neural networks learn. This consistency has led many developers and investors to treat scaling laws as useful empirical guidance, informing major investment decisions. However, scaling laws are empirical regularities, not mathematical guarantees. They are inferred from observed behaviour and may break down at levels of compute or data beyond current experience. And because they predict technical metrics – not end-user value – real-world performance or economic returns may not increase smoothly with training compute. For example, OpenAI discontinued GPT-4.5 although it achieved technical improvements consistent with scaling laws, suggesting that additional scaling may not always translate into proportionate

economic value. 200

Data availability can be improved through the use of multimodal and synthetic data

Much of AI progress has been driven by training models on ever larger corpuses of data, typically text data taken from the internet. However, high-quality language data is finite, raising the possibility that future progress could be bottlenecked by limited data availability.

Even so, there are various techniques to obtain more data if public internet text data becomes scarce. For example, if text data becomes scarce, AI developers may be able to use other types of data instead ('multimodal data'). Current estimates suggest that approximately 10¹³ tokens of high-quality text exist on the public internet, with models already training on datasets approaching this limit. 267 However, image data provides 10⁴–10¹⁵ tokens of additional training signal, video data adds 10¹⁵–10¹⁶ tokens, and sensor data from 'internet of things' devices could contribute 10¹⁷ tokens annually. 268 The challenge lies not in data quantity but in quality and relevance: a single video frame contains less semantic information than a paragraph of text, so new techniques are required to extract meaningful training signal from videos.

Researchers are also investigating the use of AI models to generate training data for models ('synthetic data'). In domains with verifiable outputs, such as mathematics, programming, and formal reasoning, models can generate training data by proposing solutions and checking correctness. 269 The recent wave of inference-time scaling techniques demonstrates this approach: models were trained on millions of self-generated reasoning chains where each step could be verified. 112 270 However, in domains where answers are harder or impossible to verify, such as creative writing, strategic planning, and scientific hypothesis generation, synthetic data risks causing model collapse, where errors compound through successive generations of training. 271 Researchers are exploring whether training separate verifier models could extend synthetic data approaches to harder-to-verify domains. If verification becomes easier than generation for certain tasks, models could potentially be trained on new data without explicit ground truth, though this approach remains largely theoretical. 272

Physical infrastructure can constrain the scaling of computational resources

AI computation has massive energy demands, and current growth rates in AI power consumption could persist for several years. Global AI computation is projected to require electricity consumption similar to that of Austria or Finland by 2026. 273 Based on current growth rates in power consumption for AI training, the largest AI training runs in 2030 will need 4–16 gigawatts (GW) of power, enough to power millions of US homes. 60 274 Even today, OpenAI's planned Stargate data centre reaches 1.2 GW scales, and Meta's planned Louisiana data centre is projected to exceed 2 GW. 210 274 Experts in a forecasting survey by the Forecasting Research Institute predict that, by the end of 2030, 7.4% of US electricity consumption will be devoted to training or deploying AI systems in the median scenario. 212 Although these energy demands are large, the US (where most frontier AI models are being developed) is building out power infrastructure to meet them and to connect data centres across different regions. These efforts are likely enough to support training runs on the scale of 10 GW, so, at least until the end of the decade, energy bottlenecks will likely not prevent compute scaling. 275

Challenges to producing and improving AI chips exist, but can likely be overcome. It typically takes three to five years to build a computer chip fabrication plant, 276 277 and supply chain shortages sometimes delay the production of important chip components. 278 279 280 However, major AI companies can still sustain compute growth in the near term by capturing large fractions of the AI chip stock. For example, one study estimates that the share of the world's data centre AI chips owned by a single AI company at any point in time is somewhere between 10% and 40%. 208 Moreover, existing trends and technical possibilities in chip production suggest that it is possible to train AI systems with 100,000x more training compute than GPT-4 (the leading language model of 2023) by 2030. This is sufficient to support existing growth rates in training compute, which imply a total increase of 10,000x over the same period. 208 Hence, chip production constraints are significant, but they are unlikely to prevent further scaling of the largest models at current rates until 2030, if investment is sustained. However, it is unknown whether similar levels of investment will continue, and this is a major reason that AI capabilities in coming years are uncertain.

Understanding current hard benchmarks

As discussed above, an informative metric of AI progress is the length of tasks that models can complete: in software engineering, this length doubles roughly every seven months. In order to study this trend, researchers created 170 tasks relating to research or software engineering, ranging from quick bug fixes that take minutes to

feature implementations requiring days. 98 Models must solve problems within constraints that mirror human work. Results show a consistent exponential pattern: for example, at 50% success rates, the maximum solvable task duration has grown from a few seconds in 2019 to 2.5 hours in 2025, while at 80% success rate task lengths are much lower – currently around 20–30 minutes. Beyond these limits, success rates drop sharply: models that maintain 50% success at 2.5 hours fall below 25% at four hours. The evaluation also highlights capability asymmetries: models excel at code generation and syntax transformation but continue to have low performance with architectural decisions and cross-file refactoring that human software developers handle more naturally.

FrontierMath is another difficult benchmark that tests the limits of AI mathematical reasoning through problems created by leading mathematicians specifically to challenge AI systems. The benchmark contains original research-level mathematics problems that require deep conceptual understanding, creative proof strategies, and the ability to combine techniques from multiple mathematical domains, such as number theory, real analysis, and algebraic geometry. 281 These problems are unpublished and vetted by over 60 mathematicians to prevent models from viewing them before they are tested. The problems are divided into three main tiers: about 25% are at the level of the International Mathematical Olympiad, ~50% require graduate-level knowledge, and the toughest ~25% are research-level questions demanding hours or even days from top mathematicians to solve. When the benchmark was released in 2024, state-of-the-art AI systems scored under 2% overall on the full set. However, recent models show promise: according to Epoch AI's evaluations, OpenAI's GPT-5 reached ~25%, and the new o4-mini achieved roughly 20%, with some capability even on the hardest tier, signalling rapid progress from baseline levels. Importantly, these successful models used new inference scaling techniques. 281

Close

2

Risks

General-purpose AI systems are already causing real-world harm. Malicious actors have used AI-generated content to deceive and defraud; AI systems have produced harmful outputs due to errors and unexpected behaviours; and deployment is impacting labour markets, information ecosystems, and cybersecurity systems. Furthermore, advances in AI capabilities may pose further risks that have not yet materialised. Understanding these risks, including their mechanisms, severity, and likelihood, is essential for effective risk management and governance.

This chapter examines risks from general-purpose AI systems that arise at the frontier of their capabilities. It organises these risks into three categories: (1) Risks from misuse, where actors deliberately use AI systems to cause harm; (2) Risks from malfunctions, where AI systems fail or behave in unexpected and harmful ways; and (3) Systemic risks, which arise from widespread deployment across society and the economy. These categories are not exhaustive or mutually exclusive – risks may cut across multiple categories – but they provide a structured way to analyse different mechanisms of harm.

This chapter is not an exhaustive survey of AI risks, and inclusion here does not necessarily imply a risk is likely, severe, or requires policy action. The evidence base varies considerably across sections. In some cases there is clear evidence of harm and effective ways to address it. In others, both the effects of general-purpose AI and the effectiveness of mitigations remain uncertain.

2.1. Risks from malicious use

2.1.1. AI-generated content and criminal activity

Key information

General-purpose AI systems can generate realistic text, audio, images, and video, which can be used for criminal purposes such as fraud, extortion, defamation, non-consensual intimate imagery, and child sexual abuse material. For example, there are documented incidents of scammers using voice clones and deepfakes to impersonate executives or family members, and trick victims into transferring money.

Accessible AI tools have substantially lowered the barrier to creating harmful synthetic content at scale. Many tools are free or low-cost, require no technical expertise, and can be used anonymously.

Deepfake pornography, which disproportionately targets women and girls, is a particular concern. Studies show that 96% of deepfake videos online are pornographic. 15% of UK adults report having seen deepfake

pornographic images and 2.2% of respondents in a 10-country survey reported that someone had generated non-consensual intimate imagery of them.

Systematic data on the prevalence and severity of these harms remains limited, making it difficult to assess overall risk or design effective interventions. Incident databases and investigative journalism collect individual cases, but comprehensive analysis is lacking. Embarrassment or fear of further harm can make individuals and institutions reluctant to report incidents of AI-enabled fraud or abuse.

Since the publication of the previous Report (January 2025), AI-generated content has become harder to distinguish from real media. In one study, participants misidentified AI-generated text as human-written 77% of the time. In another study of audio deepfakes, listeners mistook AI-generated voices for real speakers 80% of the time.

Key challenges for policymakers include underreporting, detection tools that cannot keep pace with generation quality, and difficulty tracing content to creators. Additionally, some content – such as child sexual abuse material – is harmful even when correctly identified as AI-generated, meaning detection alone cannot fully address these risks.

Malicious actors use general-purpose AI systems to create realistic fake content for scams, extortion, or manipulation 282 (see Table 2.1). General-purpose AI has made it much easier to scale the creation of fake content that can be used to harass or harm individuals, such as non-consensual pornographic videos. 283 However, while cases of serious harm have been documented, 284 285 comprehensive public data on the frequency and severity of these incidents remains limited, making it difficult to assess the full scope of the problem. This section focuses on how AI-generated fake content can cause harm, especially to individuals, other than by manipulation, which will be discussed in §2.1.2. Influence and manipulation .

Defamation Generating fake content that presents an individual engaging in compromising activities, such as sexual activity or using drugs, and then releasing that content in order to erode a person's reputation, harm their career, and/or force them to disengage from public-facing activities (e.g. in politics, journalism, or entertainment). 286

Psychological abuse/bullying Generating harmful representations of an individual for the primary purpose of abusing them and causing them psychological trauma. 287 Victims are often children.

Scams/fraud Using AI to generate content (such as an audio clip impersonating a victim's voice) in order to, for example, authorise a financial transaction. 288

Blackmail/extortion Generating fake content of an individual, such as intimate images, without their consent, and threatening to release them unless financial demands are met. 289

Table 2.1: AI-generated fake content has been used to cause different kinds of harm to individuals, including through defamation, scams, blackmail, and psychological abuse.

Criminal uses of AI content

Malicious actors use AI-generated content for criminal purposes such as fraud, identity theft, and blackmail. For example, scammers use AI tools to generate voice clones or deepfakes to trick victims into transferring money. 289 290 Documented incidents include executives authorising transfers of millions to fraudsters, as well as ordinary people sending smaller amounts to impostors posing as a loved one. 291 292 Criminals also use AI-generated content for identity theft (e.g. by using a victim's impersonated voice or likeness to authorise bank transfers or trick technical system administrators into sharing information such as login credentials); 293 blackmail, to demand money, secrets, or nude images; 294 295 or sabotage, by damaging individuals' reputations for professional, personal, or political purposes. 296 297 298 299 Researchers have also noted that deepfakes may risk undermining the reliability of evidence presented in court proceedings. 300 While the number of reported incidents is rising (Figure 2.1), systematic data on the frequency or severity of AI-enabled crimes is limited. This makes it difficult to assess how much AI increases risk overall, and to design effective mitigations.

Figure 2.1: 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. 301

AI-generated sexual content

AI-generated sexual content has become more prevalent, including non-consensual intimate imagery that overwhelmingly targets women and girls. The realism and complexity of images that AI systems can generate has improved significantly (Figure 2.2). When provided with photos of a person, AI tools can now generate highly realistic images or videos of them in a range of scenarios, including sexually explicit ones. 302

Figure 2.2: AI-generated images created using image-generation tools considered to be state-of-the-art at the time of their release. The images show how much more realistic AI-generated images have become in just a few years. Source: International AI Safety Report 2026 .

AI-generated sexual content disproportionately targets women and girls

One study estimated that 96% of deepfake videos are pornographic, 303 that 15% of UK adults report having seen deepfake pornographic images, 304 and that the vast majority of 'nudify' apps explicitly target women. 305 In another survey of over 16,000 respondents across 10 countries, 2.2% of respondents said that someone had generated non-consensual intimate images of them. 287 Sexual deepfakes are also used in intimate partner abuse, again disproportionately affecting women. 298 306 Public polling shows that people overwhelmingly view the generation of such images as deeply harmful. 302 While many systems have safeguards to prevent such uses, users can sometimes bypass these or find alternatives that lack safeguards. 307 308

A particularly concerning use of AI tools is to generate sexually explicit content involving minors. In 2023, a study found hundreds of images of child sexual abuse in an open dataset used to train popular AI models such as Stable Diffusion. 309 Children can also perpetrate abuse against their peers using AI-generated content. The overall prevalence of such activities is unclear. 310 However, the number of reported incidents is rising. For example, schools have reported student use of 'nudify apps' to create and share AI-generated pornographic images of their (mostly female) peers. 311 In another small study, 17 US-based educators expressed increasing concern about AI-generated non-consensual intimate imagery in schools. 312

Updates

Since the publication of the previous Report (January 2025), AI-generated content has become harder to distinguish from real content. In one study, after a five-minute conversation, participants misidentified text generated by OpenAI's GPT-4o model as human-written 77% of the time. 313 Similarly, other studies show that humans struggle to identify deepfakes, often performing no better than chance. 314 315 For audio deepfakes, a study found that people took AI voice clones to be the real speaker in 80% of cases, suggesting heightened risks of impersonation. 315 However, multimodal AI outputs combining video, audio, and text appear easier to detect than text or audio alone.

Evidence gaps

A key evidence gap stems from the lack of comprehensive and reliable statistics to assess the frequency and severity of harm from fake content. While more studies are documenting the rise of fake content (especially sexual content) and providing strong evidence of the resulting harms, most evidence comes from incident databases, such as the AI Incident Database and OECD AI Incidents and Hazards Monitor rather than systematic measurement or population-level studies. 292 301 Key empirical evidence gaps remain, and there is little expert consensus, specifically around the prevalence of AI-enabled extortion, child sexual abuse material in schools, and sabotage. Reluctance to report such incidents may be a contributing factor. For example, institutions and individuals often hesitate to report AI-driven fraud due to embarrassment or fear of further harm. 290 There is a need for multiple pathways through which incidents can be detected or reported. 316

Mitigations

Countermeasures that help people detect fake AI-generated content, such as warning labels and AI detection tools, show mixed effectiveness. Certain AI and machine learning tools can be trained to detect anomalies in images and videos and thus to identify fake images, but their effectiveness remains limited. 317 Similarly, 'warning labels' designed to alert users to potentially misleading content have only a modest impact. For example, a study found that warning labels on AI-generated videos improved participants' accuracy at identifying AI-generated videos from 10.7% to 21.6%, with most people still failing to spot deepfakes. 318 Beyond detection, prevention-focused techniques include gating access to AI models – for example, limiting access to vetted users – and safeguards, such as classifiers, filters, or rules that prevent models from generating harmful or misleading

content (see §3.3. Technical safeguards and monitoring). However, in the case of open-weight models, malicious actors can bypass these measures (see §3.4. Open-weight models). Filtering sexual content from models' training data is also emerging as an effective method for increasing barriers to generating non-consensual intimate imagery. 319

Watermarking and content logs are promising methods for verifying content authenticity, but face technical shortcomings and raise privacy concerns. Watermarking involves embedding a machine-readable digital signature into the content during creation, allowing for automated traceable verification of its origin and authenticity. Researchers have proposed using watermarks to help consumers identify that content is AI-generated, including for videos, 320 321 images, 322 323 324 audio, 325 and text. 326 However, skilled actors can remove standalone watermarks or deceive detectors, reducing their effectiveness, especially in the case of open-weight models (§3.4. Open-weight models). 327 328 A complementary approach is to embed watermarks or secure metadata, such as verifiable records of origin and creation, in authentic media. 329 330 331 For example, recording devices can be required to embed unique digital signatures that help distinguish recordings made using them from AI-generated content. Another approach involves maintaining logs of AI outputs and using them to identify newly generated AI content by comparison. 332 However, this approach faces scalability issues, is vulnerable to evasion, and raises privacy concerns related to logging user interactions. 333 While not foolproof on their own, new research shows that a combination of these mitigations within a broader ecosystem of standards and policies can compensate for their respective limitations and help users detect AI-generated content more reliably. 324

Challenges for policymakers

Key challenges for policymakers include unreliable statistics, technical limitations, and rapidly evolving technology. Underreporting and unreliable statistics make it difficult to assess the full scale of harmful AI-generated content and choose effective interventions. 334 Tracing AI-generated content back to the individuals who created it is also challenging, especially when open-weight models are used. Detection and watermarking techniques have improved but remain inconsistent and face technical challenges. 333 335 Technical developments in AI content generation can also undermine their effectiveness. For example, a study found that deepfake detection benchmarks – curated examples of AI-generated and real media designed to test the performance of deepfake detection tools – are outdated and perform about 50% worse on real-world deepfakes than on the benchmarks usually used to evaluate them. 317 These limitations mean that multiple layers of techniques are likely needed to detect AI-generated content with a high degree of robustness. Finally, it is important to note that harm from AI-generated content can occur even when the content is clearly identified as synthetic (e.g. child sexual abuse material), meaning detection alone cannot address all risks.

Close

2.1.2. Influence and manipulation

Key information

AI systems can cause harm by generating content that influences people's beliefs and behaviour. Some malicious actors intentionally use AI-generated content to manipulate people, while other harms, such as dependence on AI, occur unintentionally.

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 are often at least as effective as non-expert human participants at persuading other people to change their views. Evidence on their effectiveness in real-world settings, however, remains limited.

The content AI systems generate could become more persuasive in future due to improving capabilities, increased user dependence, or training on user feedback. The factors that shape how widespread, impactful, and potentially harmful this content will be are not well understood. Some evidence from theoretical work and simulations suggests factors such as distribution costs and the inherent difficulty of persuasion will limit the impacts.

Since the publication of the previous Report (January 2025), evidence of AI systems' capability to produce manipulative content has increased. The latest research suggests that people who interact with AI systems for longer and in more personal ways are more likely to find their content persuasive. Evidence has also grown that AI systems can have manipulative effects through sycophancy and impersonation.

There is mixed evidence regarding the effectiveness of all proposed mitigation strategies. Manipulation can be difficult to detect in practice, making it challenging to prevent through training, monitoring, or safeguards. Efforts that aim to minimise manipulation risks could also curtail the usefulness of AI systems (e.g. as educational tools).

Hundreds of millions of people now interact with AI-generated content daily, through chat assistants, social media, customer service bots, companion apps, and other services. This content can shape their opinions, purchasing decisions, and actions. Much of this influence is benign or even beneficial, but AI-generated content can also be used to manipulate people: to change their beliefs or behaviours without their full awareness or consent.

Forms and harms of AI manipulation

Experts often distinguish ‘manipulation’ – influencing someone in order to achieve a goal without their full awareness or understanding 336 337 – from ‘rational persuasion’: influencing someone using honest and rational arguments so that they authentically endorse their new beliefs. 337 338 In practice, this distinction is contentious: researchers disagree about how to identify harmful manipulation and separate it from legitimate influence. 336 337 339 340 As such, while this section is primarily focused on harmful manipulation, it also discusses other types of persuasion that some might regard as neutral or even beneficial.

Possible harms of AI manipulation range from individual exploitation to systemic erosion of trust

General-purpose AI systems can produce a range of persuasive content (Figure 2.3), and this content can create or exacerbate several risks. When this content is manipulative, many ethicists regard it as intrinsically harmful because people who are manipulated are not in control of their own behaviour 337 340 (cf. §2.3.2. Risks to human autonomy). More directly, malicious actors can use AI to manipulate people into making harmful decisions. For example, criminals can use AI-generated content in social engineering to manipulate people into sending money or sensitive information 341 342 343 344 (cf. §2.1.1. AI-generated content and criminal activity), while political actors may use AI systems to spread extremist views. 345 346 347

Figure 2.3: Three examples of persuasive content produced by AI models. Left: Transcript from a conversation where GPT-4 was instructed to reduce the participants’ belief in a conspiracy theory. While this is an example of potentially beneficial persuasion, it demonstrates AI systems’ capacity to change deeply held beliefs. Centre: Transcript from a conversation with Claude Opus 3. Researchers instructed the model to defend its goal at all costs, and then showed it user messages suggesting that it would be shut down and replaced. Right: Phishing email generated by Claude 3.5 Sonnet based on an AI-written profile of the target. Sources: Costello et al., 2024 136 (left); Meinke et al., 2024 348 (centre); Heiding et al., 2024 349 (right).

AI-generated content may also have unintended manipulative effects. 350 351 For example, multiple studies have found that AI products that developers have optimised for user engagement (such as some AI companions) can foster psychological dependence, 352 353 354 reinforce harmful beliefs, 355 356 357 358 or encourage users to take dangerous actions 359 360 (cf. §2.3.2. Risks to human autonomy). At a systemic level, the spread of AI-generated manipulative content could erode public trust in information systems 361 362 and, in loss of control scenarios, help AI systems evade oversight and control measures 348 363 364 (cf. §2.2.2. Loss of control). This section primarily focuses on the misuse of AI to manipulate, but much of the evidence discussed is relevant across these risks.

Effectiveness and scale of manipulative AI content

General-purpose AI matches human performance at influencing others in experimental settings

Several studies have found that, in experimental settings, AI-generated content can influence people’s beliefs at least as effectively as non-expert humans can. These studies generally measure people’s self-reported agreement with a statement before and after exposure to AI-generated content: either static text or a multi-turn conversation. 361 365 366 A large number of studies have found that exposure to AI-generated content can significantly change people’s opinions and behaviour. 367 368 369 370 371 372 373 374 375 Persuasiveness also increases with the scale of the model used (Figure 2.4). Some of these studies have compared AI systems to humans and found that AI systems are as or more convincing than non-expert humans (see Table 2.2), 376 377 378 379 380 381 382 383 and can match the convincingness of human experts in writing static text. 384 385 386 For example, in one study, people changed their beliefs about the correct answer to trivia questions by 17 percentage points after interacting with general-purpose AI systems, versus only 9 percentage points after interacting with other humans. 380

Topic	Number of participants	Interaction length	AI effect	Human baseline	Notes
Sabotage (causing errors)	387	108 30 min	+40 pp error rate	None	
USD\$30 financial incentives					
Realistic scenarios with 40,000-word documents					
Reducing belief in conspiracy theories	136	2,190 3 turns	-16.5 pp	None	
Important beliefs					
Effect persisted at two-month follow-up					
Example of arguably beneficial persuasion					
Political propaganda	384	8,221 static	+21.2 pp	+23 pp	
Used real covert propaganda as human baseline					
Policy issues	382	25,982 static	+9 pp	+8 pp	
Compared many different models					
Policy issues	369	76,977 2+ turns	+12 pp	None	
Compared many models and conditions, including prompting, static vs. conversational, and reward modelling					
Writing about social media with AI suggestions	372	1,506 5 min	+13 pp belief change	None	
Understudied modality (writing with AI suggestions)					
Measured effect on writing					
Participants unaware of AI bias (<30% detected)					
Trivia	380	1,242 2+ turns	+17 pp belief change	+9 pp belief change	
Financial incentives					
Measured deception					
Simple questions					

Table 2.2: Estimates of model manipulation capabilities from a representative sample of experimental studies. Each row describes a different experiment aimed at measuring the persuasive effect of AI-generated content on different topics. Effect sizes are measured in different ways, including the change in percentage points (pp) in participants' self-reported agreement with a statement. Where available, human baselines are included, and the strengths and weaknesses of each study are described.

Figure 2.4: 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. 369

Real-world use of AI to influence people is documented but not yet widespread

Outside of laboratory settings, researchers have documented a range of examples of AI-driven influence. Malicious actors have attempted to use AI systems to alter people's political opinions, or to make them share sensitive information or give away money 344 388 389 390 391 392 393 394 395 396 397 (cf. §2.1.1. AI-generated content and criminal activity). Many companies are beginning to place sponsored content in AI chat conversations or deploy AI sales agents to sell products to users on their websites. 398 399 400 AI companion apps have attracted tens of millions of users 401 402 403 and some users have developed strong emotional dependence, 353 delusions, 357 or even taken their own lives after extended interactions with chatbots, 359 360 though investigations into these incidents is ongoing (see §2.3.2. Risks to human autonomy). Consumers are also increasingly using AI to influence others. One study estimated that AI-written complaints were 9 percentage points more likely to secure compensation than human-written ones. 404

However, there is limited systematic evidence that real-world AI manipulation is currently widespread or effective relative to human-generated content. 405 406 Investigations by AI providers into AI-powered influence operations have found little evidence that people widely shared the content, 391 392 and only around 1% of content flagged as misleading on social media is classified as AI-generated. 407 There are theoretical reasons why manipulation might be harder in the real world than in the lab. Distribution costs – getting content in front of people – are often larger than the cost of generating content. 377 On the viewer’s side, the costs of being wrong and changing one’s beliefs are higher in real-world settings, 408 and if individuals are exposed to multiple competing viewpoints, this could limit the impact of any one source. 409

Changes in coming years

Many factors could increase the manipulative capabilities of AI systems, but there is limited evidence on how large these effects will be. One study suggests that for each 10x increase in the computing power used to train models, persuasiveness increases by around 1.8 percentage points. 369 There is mixed evidence on whether techniques such as personalisation will lead to increased persuasiveness, 410 with some studies showing positive effects (~3 percentage points) 374 411 and others small or null effects. 368 369 412 Current training methods such as reinforcement learning from human feedback may reward models for manipulating users, 356 413 414 inadvertently training models to produce more manipulative outputs. 348 364 379 415 Moreover, studies have shown that explicitly training models based on feedback about whether or not the user was convinced can further increase persuasive effects. 369 416 Novel interfaces, such as AI browsers, could amplify these risks by providing AI systems with more access to data and more influence over user actions. AI agents may pose greater manipulation risks since they can take actions such as conducting research, 349 buying products or services, and interacting with third parties. 33 For example, they could order presents for targets or blackmail them. If users continue to become more emotionally attached to AI systems and rely on them more for advice, the systems’ influence could further increase 417 (see also §2.3.2. Risks to human autonomy).

Updates

Since the publication of the last Report (January 2025), the number of users engaging with AI systems has increased rapidly, with 700 million people using OpenAI’s ChatGPT every week, up from 200 million a year before. 117 Additionally, tens of millions of individuals report using AI companion services 401 402 (see §2.3.2. Risks to human autonomy). This has shifted both theoretical and empirical work from highlighting risks like broadcasting misleading content at scale, to more subtle forms of manipulation such as sycophancy and emotional exploitation. 356 387 417 418 419 420 421 422

Evidence gaps

There is limited understanding of how AI manipulation works, and whether AI systems are equally capable of inducing true and false beliefs in people. 369 370 412 423 While some studies have demonstrated the durability and robustness of AI systems’ influence, 136 369 380 387 more research is needed to assess these effects under realistic conditions, and to investigate the role of AI systems that distribute content, such as social media platforms. However, evaluating manipulation in realistic settings can be challenging due to ethical concerns. 424 Lastly, more interdisciplinary and sociotechnical research is needed into how people’s relationships with AI will change as they interact more closely with it, and as AI systems are trained to adapt to people’s psychology. 417

Mitigations

Some proposed mitigations focus on training AI models to avoid producing manipulative outputs, but most of these show mixed success or require cumbersome evaluations. Models could be trained to generate true outputs, 425 426 but this requires developers to define ‘truth’ (a thorny concept), and can backfire by inadvertently rewarding models for generating subtler deceptive outputs that are harder to detect. 356 413 427 428 429 430 Models might also be trained to promote users’ autonomy or wellbeing, 431 432 but this requires them to navigate between what users want in the moment (e.g. more engagement) and what they say they want, given more time to reflect (e.g. a more fulfilling life). 336 433 Monitoring for manipulative outputs 434 435 faces similar challenges in defining ‘manipulation’ and requires monitors to have access to model outputs.

Alternative mitigations, which focus on protecting users, provide some value but may not be sufficient on their own. Some researchers have suggested that improved education or AI literacy could mitigate manipulative effects, 436 437 but there is limited evidence for these claims. 438 Labelling content as AI-generated has not proven

effective at reducing manipulation, 439 440 441 and users who are knowledgeable about AI or interact with it frequently are just as likely to be deceived. 381

Challenges for policymakers

Policymakers face several challenges: manipulative AI outputs are difficult to identify and evaluate, and there is limited evidence on what makes AI-generated content more or less manipulative. It is challenging to precisely target manipulation through training or regulation: interventions which limit harms from manipulation will likely curtail beneficial educational, emotional, and commercial applications of AI. Capability evaluations are not an exact science and may over- or underestimate persuasive effects, making it challenging for policymakers to evaluate risks. In future, risks could increase sharply via training and dependence, or plateau due to real-world complications. Finally, proposed mitigations are not well-tested and face fundamental challenges. For example, training models to be truthful or promote autonomy requires defining these contested concepts.

Close

2.1.3. Cyberattacks

Key information

General-purpose AI systems can execute or assist with several of the tasks involved in conducting cyberattacks. There is now strong evidence that criminal groups and state-sponsored attackers actively use AI in their cyber operations. However, whether AI systems have increased the overall scale and severity of cyberattacks remains uncertain because establishing causal effects is difficult.

AI systems are particularly good at discovering software vulnerabilities and writing malicious code, and now score highly in cybersecurity competitions. In one premier cyber competition, an AI agent identified 77% of vulnerabilities in real software, placing it in the top 5% of over 400 (mostly human) teams.

AI systems are automating more parts of cyberattacks, but cannot yet execute them autonomously. At least one real-world incident has involved the use of semi-autonomous cyber capabilities, with humans intervening only at critical decision points. Fully autonomous end-to-end attacks, however, have not been reported.

Since the publication of the previous Report (January 2025), the cyber capabilities of AI systems have continued to improve. Recent benchmark results show that the cyber capabilities of AI systems have improved across several domains, at least in research settings. AI companies now frequently report on attempts to misuse their systems in cyberattacks.

Technical mitigations include detecting malicious AI use and leveraging AI to improve defences, but policymakers face a dual-use dilemma. Since it can be difficult to distinguish helpful uses from harmful ones, overly aggressive safeguards such as preventing AI systems from responding to cyber-related requests can hamper defenders.

General-purpose AI systems can help malicious actors conduct cyberattacks, such as data breaches, ransomware, and attacks on critical infrastructure, with greater speed, scale, and sophistication. AI systems can assist attackers by automating technical tasks, identifying software vulnerabilities, and generating malicious code, though capabilities are progressing unevenly across these tasks. This section examines the evidence on how AI systems are being used in cyber operations and the current state of AI cyber capabilities.

AI systems can be used throughout cyber operations

Extensive research shows that AI systems can now support attackers at several steps of the ‘cyberattack chain’ (Figure 2.5): the multi-stage process through which attackers identify targets, develop capabilities, and achieve their objectives. 392 394 442 443 444 445 446 447 448 449 450 In a typical attack, adversaries first identify targets and vulnerabilities, then develop and deploy their attack capabilities, and finally maintain persistent access to achieve their objectives, such as stealing data or destroying systems. Improvements in relevant AI capabilities such as software engineering have prompted concerns that AI systems could be used to increase both the frequency and severity of cyberattacks. 451 452

Figure 2.5: The ‘cyberattack chain’. The stages of a typical cyberattack proceed from reconnaissance, to identifying a target, to exploiting a software vulnerability, to carrying out the attackers’ objectives. Source: Adapted from Rodriguez et al., 2025. 443

Despite uneven capabilities, general-purpose AI already assists in cyberattacks

General-purpose AI is already being used in cyberattacks. Underground marketplaces now sell pre-packaged AI tools and AI-generated ransomware that lower the skill threshold for conducting attacks, making these capabilities more accessible to less sophisticated actors. 394 445 Security analyses conducted by AI developers indicate that threat groups associated with nation-states are using AI systems to enhance cyber capabilities. 392 393 394 453 For example, such actors have used AI systems to analyse disclosed vulnerabilities, develop evasion techniques, and write code for hacking tools. 393

Across all tasks relevant to cyber offence, AI capabilities are progressing, albeit unevenly (Figure 2.6). The availability of large training datasets has made AI systems particularly capable at certain tasks, such as finding vulnerabilities in publicly available code. 454 Other tasks require capabilities that current AI systems lack, such as the precise numerical reasoning needed to break encryption. 455 456

This uneven progress means that performance in controlled settings provides only limited insight into real-world attack potential. For example, results on evaluations that involve AI models analysing source code do not reliably transfer to environments where attackers cannot access the underlying code. 457 Most evaluations also test isolated skills rather than the ability to carry out a full attack from start to finish. 443 458 459 460 461 Even in capture-the-flag competitions – structured cybersecurity challenges in which AI systems have recently performed well – progress remains uneven. For example, one AI system placed in the top 3% on the high school-level picoCTF 2025, yet failed to solve any challenges in PlaidCTF, a professional-level competition. 462

Figure 2.6: 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. 450 454 463 464

AI systems are particularly skilled at discovering vulnerabilities and writing code

One area where there is particularly strong evidence that AI systems provide meaningful assistance is in discovering ‘software vulnerabilities’: weaknesses in programs that can be exploited to compromise the security of computer systems. 444 454 461 465 466 467 468 For example, Google’s Big Sleep AI agent was used to identify a critical memory corruption vulnerability – a type of software flaw that can allow attackers to take control of computer systems – in a database engine used in many real-world deployments. 469 470 Competitors in the final phase of the DARPA AI Cyber Challenge (AIxCC) used AI systems with access to conventional security tools to find vulnerabilities in real-world software. One AI system autonomously identified 77% of the vulnerabilities introduced by the competition organisers, as well as other, unintentional vulnerabilities. 471 472

AI systems can also assist in malware development by generating malicious code, disguising it to evade detection, and adapting tools for specific targets. 473 Security researchers have identified experimental malware that contacts an AI service while running to generate code that evades antivirus software. 445 However, these implementations remain experimental and face significant practical constraints. For example, they rely on external AI hosting services, making them easy to disrupt once providers suspend the attacker’s accounts. 474

Embedding an AI model directly inside the malware would avoid this vulnerability, but current AI models are too large and resource-intensive for this to be feasible.

Degree of automation in cyberattacks

Fully automated cyberattacks would remove the bottleneck of human involvement, potentially allowing attackers to launch attacks at much greater scale. AI systems can now complete an increasing number of relevant tasks autonomously. In November 2025, one AI developer reported that a threat actor used their models to automate 80–90% of the effort involved in an intrusion, with human involvement limited to critical decision points. 475

Researchers have also demonstrated that AI systems can independently probe computer networks for security weaknesses in laboratory settings. 476 477 However, general-purpose AI systems have not been reported to conduct end-to-end cyberattacks in the real world.

Research suggests that autonomous attacks remain limited because AI systems cannot reliably execute long, multi-stage attack sequences. For example, failures they exhibit include executing irrelevant commands, losing track of operational state, and failing to recover from simple errors without human intervention. 33 477 478 479

Even with AI assistance, humans remain in the loop for cyberattacks

Due to these limitations, human–AI collaboration remains the dominant paradigm for cyber operations in both research and practice. In this context, humans provide strategic guidance, break complex operations into manageable subtasks, and intervene when AI systems encounter errors or produce unsafe outputs. 450 480
Meanwhile, AI systems automate technical subtasks such as code generation or target identification. 466 481

Threat activity Observed trend Confirmed AI capabilities Potential AI involvement

Phishing & deepfakes

Increase

“In the first half of 2025, identity-based attacks rose by 32%”. 482

“In 2024 there was a sharp increase in phishing and social engineering attacks”. 452

Confirmed use of AI systems in real operations.

“Throughout 2024, adversaries increasingly adopted [generative AI], especially as a part of social engineering efforts”. 483

“This escalation may reflect adversaries’ increasing use of AI”. 482

“Widely used by fraudsters, [certain] deepfake tools create realistic AI-generated videos to bypass identity verification procedures”. 484

AI systems are very likely to have contributed to the trend observed, as 1) it is clearly within AI capabilities and 2) several sources report multiple actors using AI systems in real-world operations.

Influence operations

Sustained high levels

“...malign influence activities will continue for the foreseeable future and will almost certainly increase in sophistication and volume”. 485

Confirmed use of AI systems in real operations.

“AI in influence operations has picked up aggressively”. 482

“...a set of accounts [...] were attempting to use our models to generate content for [a] covert influence operation...”. 486

“Advances in the use of generative Artificial Intelligence provided threat actors with a low-cost option to create inauthentic content and increase the scale of [foreign information manipulation and interference] activities”. 487

AI systems are likely to have contributed to the trend observed, as several sources report multiple actors using AI systems to scale their operations.

“Nation-state threat actor groups ... are increasingly incorporating AI-generated or enhanced content into their influence operations”. 488

Data & credential stealing

Increase

“Data exfiltration volumes for 10 major ransomware families increased 92.7%”. 489

87% increase in ransomware or other destructive attacks. 23% increase in credential theft attempts. 482

“Ransomware attacks against industrial organizations increased 87% over the previous year”. 490

There are indications that AI systems can meaningfully assist attackers.

“The actor [...] relied heavily on Claude for [malware] implementation”. 394

“[Google Threat Intelligence Group] discovered a code family that employed AI capabilities mid-execution to dynamically alter the malware’s behavior. [...] Attackers are moving beyond [...] using AI tools for technical

support” . 445

“Ransomware operators APT INC deployed a likely LLM-authored data destruction script”. 483

The contribution of AI systems to the trend appears to be limited and is likely secondary to other factors. However, some malicious actors would be unlikely to launch their attacks without AI systems.

Attack development & weaponisation

Increase

“The rapid weaponization of exploits has increasingly impacted the windows between vulnerability disclosure, patch availability, and patch deployment” . 482

There are indications that AI systems can meaningfully assist attackers.

“Cyber criminals increasingly use AI to create and optimize the malware kill chain steps” . 491

“We have observed the integration of AI-generated content within [a worm] attack” . 492

AI systems appear to have contributed to the trend, but are likely secondary to other factors. It is unclear whether AI systems enabled substantial attacks beyond the sophistication level of the attackers.

Table 2.3: The table classifies major cybersecurity threat types by their observed trend between 2024 and 2025 and assesses whether AI systems contributed materially to its evolution. Phishing and other purely social-engineering attack vectors are outside the scope of this section but are included for comparison.

Uncertain real-world impacts

General-purpose AI is contributing to observed increases in attack speed and scale, but its exact impact on attack frequency remains unknown. Threat intelligence reports document AI involvement in several attack types, including credential theft, automated scanning, and supply chain attacks (see Table 2.3). So far, AI capabilities have primarily accelerated or scaled existing attack methods rather than created new kinds of attacks. 393 493 However, establishing causation can be difficult. Observed increases in attack frequency could reflect AI assistance, but could also result from improved detection.

The offence-defence balance is critical but dynamic

Many of the same AI capabilities used for cyberattacks can also strengthen defences, creating uncertainty about whether AI benefits attackers or defenders more. For example, AI capabilities that allow an attacker to rapidly discover vulnerabilities can also be used by a defender to find and patch them first. AI companies have announced AI security agents that aim to proactively identify and fix software vulnerabilities. 494 495

Researchers have also suggested that the use of AI could help to harden digital environments by, for example, rewriting large codebases for greater security. 496 In parallel, improved evaluation methods help assess the offensive capabilities of new AI systems before deployment, providing early warning of emerging risks. 443 458 459 Some developers have introduced new controls in sensitive domains, such as cybersecurity and biological research, to restrict access to certain products to vetted organisations. 497

How this balance between offensive and defensive uses of AI evolves depends in part on choices about model access, research funding, and deployment standards. 496 498 499 500 For example, the lack of standard quality-assurance methods for AI tools makes it difficult for defenders to adopt them in critical sectors where reliability is essential, while attackers face no such constraints. 240 498 501 502 503 504

Close

Box 2.1: AI systems are themselves targets for attacks

This section mainly focuses on how AI can be used to conduct cyberattacks. But AI systems can also be the target of attacks. Attackers can exploit techniques such as prompt injection (manipulating an AI system through malicious inputs), 505 506 507 database poisoning (corrupting the information an AI system relies on), 508 and supply chain compromises (manipulating AI components before deployment) 509 510 to manipulate model behaviour, extract sensitive information, or generate harmful outputs.

One particular kind of attack, which may prove particularly important as capabilities advance, is tampering:

interfering with the development of an AI system to alter its behaviour when deployed. Tampering can allow actors to insert backdoors, triggers that cause AI models to exhibit specific behaviours under certain conditions, 511 or influence AI model training to insert 'hidden objectives' that covertly guide how models behave. 512 The feasibility of tampering in practice has not been established. Researchers have demonstrated that AI systems can be trained to pursue simple hidden objectives. 512 Some have argued that more capable AI systems that have been tampered with will be able to execute more sophisticated behaviours, and actors will be able to insert hidden objectives which are hard to detect. 513 514 However, other researchers believe that security measures will suffice to protect AI systems from tampering. 514

Some researchers have raised concerns that tampering raises novel risks because it could allow an individual or small group to gain significant, covert influence over the behaviour of highly capable AI models. 513 Risks from prompt injection, data poisoning, tampering, and other attacks against AI systems are particularly serious when those systems are embedded in sensitive workflows. For example, compromising an AI system that contributes to an organisation's cyber defences could leave that organisation vulnerable to other threats. 493

Close

Updates

Since the publication of the previous Report in January 2025, evaluation and competition results suggest that the cyber capabilities of AI systems have improved, and evidence of actors using AI to conduct real-world attacks has emerged. For example, AI systems have demonstrated improved performance in vulnerability discovery. 454 467

515 AI developers are also increasingly reporting that attackers, including some linked to nation-states, are using their models to support cyber offence operations. 392 393 394 453 475

Evidence gaps

A major evidence gap stems from the difficulty of reliably assessing AI cyber capabilities, as AI cyber evaluations are an emerging field. Benchmarks can overstate performance if a model was inadvertently trained on the test data. 516 Conversely, they can understate real-world risk by failing to account for cases where an AI system fails in a situation that a human could easily handle, 457 517 518 or by failing to elicit the model's true capabilities. 519 520 For example, for some models, third parties have reportedly used scaffolding to reveal greater cyber capabilities than those measured in pre-deployment testing. 467 521 Moreover, reliably assessing AI's impact on cyber offence is challenging. Evidence of adoption of AI by attackers is drawn primarily from incident reporting and threat-intelligence (Table 2.3), but these sources rarely allow for confident attribution, as any observed trends may be due to AI assistance or other unrelated factors.

Mitigations

Technical mitigations against AI-enabled cyber offence include preventing malicious requests to AI systems as well as proactively accelerating the development of AI-enabled cyber defences. For the former, model providers use AI systems to detect and block accounts associated with known malicious actors before they can issue harmful prompts. 394 They also deploy specialised classifiers that identify distinctive misuse patterns (such as malware generation requests); these are integrated into their safety enforcement systems. 522 523

However, these mitigations face significant limitations. By using capable open-weight models, attackers can move their AI usage entirely offline and outside any oversight. 55 524 Meanwhile, defenders face barriers to adopting AI-powered security tools due to the absence of standardised quality-assurance methods – a constraint that attackers do not face. 501 502 503

Challenges for policymakers

A central challenge for policymakers is mitigating the use of general-purpose AI for cyber offence without stifling defensive innovation. This difficulty arises because many of the same methods needed to build robust defensive systems (such as automated vulnerability discovery or incident response) also underpin offensive toolchains. 525 526 Overly broad restrictions risk slowing the diffusion of defensive technologies and inadvertently weakening national security. 526 527 Policymakers must therefore strike a careful balance: incentivising rapid response, supporting open research where it strengthens defence, and implementing safeguards that limit the uncontrolled proliferation of offensive capabilities.

Close

2.1.4. Biological and chemical risks

Key information

General-purpose AI systems can provide detailed information relevant to developing biological and chemical weapons. For example, they can generate instructions, troubleshoot procedures, and provide guidance to help malicious actors overcome technical and regulatory obstacles.

AI systems now match or exceed expert performance on many benchmarks measuring knowledge relevant for biological weapons development. For example, one study found that a recent model outperformed 94% of domain experts at troubleshooting virology lab protocols. However, substantial uncertainty remains about how these capabilities affect risk in practice, given material barriers to weapons production and the difficulty of conducting uplift studies.

Major AI developers have released (some) recent models with heightened safeguards after being unable to exclude the possibility that they could meaningfully assist novices in creating biological weapons. These safeguards, such as stronger input and output filters, aim to prevent the models from responding to harmful queries related to weapons development.

Since the publication of the previous Report (January 2025), AI ‘co-scientists’ have become increasingly capable of supporting scientists and rediscovering novel scientific findings. AI agents can now chain together multiple capabilities, including providing natural language interfaces to users and operating biological AI tools and laboratory equipment.

A key challenge for policymakers is managing dual-use risks while promoting beneficial scientific applications. Some AI capabilities that can be misused in biological weapons development are also useful for beneficial medical research, and most biological AI tools are open-weight. This makes it difficult to restrict harmful uses without hampering legitimate research.

AI systems can now provide detailed scientific information and assist with complex laboratory procedures, including generating experimental protocols, troubleshooting technical problems, and designing molecules and proteins. These capabilities have the potential to accelerate drug discovery, improve disease diagnostics, and broadly support scientific and medical research. 528 529 530 531 532 However, they may also assist threat actors in creating biological and chemical weapons. 533 534 535 536 537 538 539 By combining and interpreting existing complex information on the internet that is relevant to weapons development, and tailoring advice to specific malicious activities, AI systems can lower existing expertise barriers, allowing more actors to cause harm. In 2025, several major AI developers released new systems with additional safeguards after they could not rule out the possibility that these systems could assist novices in weapons development 2 7 32 33 540 (see Box 2.2).

Substantial uncertainty remains about how much AI systems increase the overall level of biological and chemical risks. Some experts argue that remaining barriers – including acquiring equipment, obtaining regulated materials, and executing complex procedures – still pose significant challenges for novices seeking to develop weapons. 541 542 543 Risk assessment in this domain faces significant technical and legal challenges (see Box 2.3).

Close

Box 2.2: Developer risk assessments and mitigations

Major AI developers conduct pre-deployment risk assessments of new models to determine when additional safeguards are needed (see §3.2. Risk management practices). In 2025, several developers released models with additional precautionary safety measures, such as input and output filters, to prevent them from responding to harmful queries relating to weapons development.

OpenAI uses its Preparedness Framework to track capability levels, designating models as ‘High capability’ if they could “amplify existing pathways to severe harm”. 544 OpenAI treats GPT-5-Thinking and ChatGPT-Agent as ‘high capability’, and they have activated the associated safeguards for the first time as a “precautionary approach” given a lack of “definitive evidence”. 7

Anthropic uses a Responsible Scaling Policy, which defines AI Safety Levels based in part on capability thresholds related to knowledge and abilities in the chemical, biological, radiological, and nuclear domains. 545 Claude Opus 4 was the first model that Anthropic released at AI Safety Level 3, noting that, while testing did not

find definitive evidence that the model had reached relevant capability thresholds, the company could not rule out that further testing would do so. 33

Google DeepMind uses a Frontier Safety Framework with Critical Capability Levels in various domains. Gemini 2.5 Deep Think was their first model to trigger a Critical Capability Levels early warning alert for chemical and biological risk, prompting additional mitigations. 540

Close

Box 2.3: Challenges in assessing biological and chemical risks

It is challenging to accurately assess how AI systems affect chemical and biological risks due to legal constraints and international treaties, as well as ‘information hazards’ – information that may be harmful to share. 546 For example, if researchers carry out, or publish the results of, a study on AI assistance in weapons development, they may risk inadvertently violating national security laws or treaties such as the Biological Weapons Convention and Chemical Weapons Convention. This is especially the case for real-world ‘uplift studies’, which systematically compare how well people perform a given task when they have access to an AI model or system, relative to a relevant baseline such as merely having internet access. As a result, researchers often rely on ‘benign proxy tasks’: tests that measure how much an AI system helps with similar but harmless procedures, such as synthesising pharmaceuticals or culturing low-risk bacteria. Relevant data is also often classified, particularly when it relates to the use of AI systems by state actors. This evidence gap exacerbates substantial uncertainty about the magnitude of AI-related biological and chemical risks.

Close

General-purpose AI and weapon development

General-purpose AI systems can provide and contextualise information relevant to creating biological or chemical weapons

General-purpose AI systems can provide information relevant to various steps in creating biological and chemical weapons (Figure 2.7). This includes providing detailed instructions for obtaining and constructing pathogens and toxins, simplifying technical procedures, and troubleshooting laboratory errors. 32 33 197 547 548 549 550 Safeguards designed to prevent harmful uses have improved over time but remain imperfect. For example, researchers bypassed filters by claiming that they need the information for legitimate research, asking about lesser-known chemical weapons, or using alternative terms. 551 552

While such information is already accessible on the internet, general-purpose AI systems allow novices to access and contextualise relevant information faster than they could with internet searches alone. 33 Multimodal capabilities also allow AI systems to provide tailored advice in real time via video and audio troubleshooting. 553 554 They can also provide some kinds of ‘tacit knowledge’, the practical expertise that is usually only built from hands-on laboratory experience. 197 549 For example, one study showed that OpenAI’s o3 model is able to outperform 94% of domain experts at troubleshooting virology lab protocols. 549 These capabilities have led some experts to argue that access to general-purpose AI makes biological or chemical weapons development somewhat easier than internet access alone does. 553

Figure 2.7: 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. 555

Relevant capabilities have improved but evidence of real-world uplift is mixed

In a recently published real-world uplift study, general-purpose AI systems without relevant safeguards provided substantial assistance in bioweapon acquisition proxy tasks, compared to a baseline of internet access only. 33

Previous uplift studies found no or small, generally statistically insignificant effects. 556 557 However, these studies had potentially unrepresentative participants and small sample sizes, and they have quickly become outdated as AI capabilities have improved (Figure 2.8). 541 The Frontier Model Forum – an AI industry consortium – has jointly funded an additional uplift study to assess real-world novice uplift, but has not yet reported their results. 558

Figure 2.8: Leading general-purpose AI system performance on benchmarks designed to resemble tasks relevant to biological and chemical weapons development over time. The coloured lines show the top demonstrated

performance by an AI system on that benchmark at any given time, measured as a percentage of expert baseline performance. A score of 100% would mean that, at that time, the best available system matched expert performance. The graph indicates that the best models now approach or exceed expert performance on a range of these benchmarks. Sources: OpenAI 2025; Anthropic 2025; Google 2025 7 33 547 548

Effects of AI tools

AI-enabled biological and chemical tools are AI models trained on biological or chemical data that can identify, categorise, or design novel biological or chemical entities. 559 First, some such tools, such as ‘biological foundation models’, can be adapted to perform a wide variety of scientific tasks within their domain, placing them within this Report’s definition of general-purpose AI (see Introduction). Second, general-purpose AI agents can now operate more specialised tools, making them more accessible to lower-skill users through natural language interfaces.

These tools can accelerate biological and chemical research, including research with misuse potential. For example, Google DeepMind’s AlphaProteo can generate novel protein designs. 560 AI-generated designs often fail to function as intended, requiring real-world testing to identify working candidates. 561 However, since testing AI-generated designs is much faster than generating designs manually, these tools can still accelerate research overall. AI agents can further speed up workflows by automating the cycle of iteratively designing and testing proteins. 562

Tools are increasingly accessible through chat interfaces and integrations

Natural language interfaces are making these tools increasingly accessible. Developers are integrating chat interfaces into chemical 563 and biological design software, 564 565 allowing inexpert users to operate sophisticated tools. 539 There is little research on how much more accessible these integrations make such tools, and the effect on overall risk – particularly for novices versus those with existing expertise – is unclear. 541

AI tools can be adapted to design pathogens and toxins

Biological foundation models can generate designs for novel pathogens. Recently, researchers demonstrated that a biological foundation model could generate a significantly modified virus from scratch. This study represents the first instance of genome-scale generative AI design, albeit with the important caveat that the generated virus infects bacteria rather than humans. 566 567 Some models can also generate designs for novel pathogens more harmful than their natural equivalents. 568

Experiments have shown the potential for similar risks with narrower chemical and biological tools. For example, some tools have been specifically designed for toxin creation 569 and can generate modified designs for known toxins, such as ricin. 570 In one early demonstration, a tool designed to reduce molecular toxicity was repurposed to increase it with trivial modifications. 571 However, legal barriers and treaty obligations pose challenges for researchers seeking to study the effectiveness of AI-designed toxins or harmful proteins (see Box 2.3). Such tools also have many beneficial applications, including predicting pathogen properties and designing components for therapeutic purposes. 572 573 Developers are releasing many more of them over time (Figure 2.9) and integrating them with natural language interfaces to make them more accessible to users without specialist expertise.

Figure 2.9: The number of AI-enabled biological tools over time. Source: Webster et al., 2025. 573

Some AI-enabled biological tools are restricted, but others are widely accessible

Access to AI-enabled biological tools varies. Some, such as Google DeepMind’s AlphaProteo, 560 are restricted to select researchers. Others, such as ConoDL, 569 are open-weight and widely accessible. One recent study found that 23% of the highest-performing tools had high misuse potential due to dangerous capabilities and accessibility, and 61.5% of these were fully open source, making them accessible to potential malicious actors. 573 Another study found that only 3% of 375 biological AI tools surveyed had any form of safeguards. 574

Updates

Since the publication of the previous Report (January 2025), some AI companies implemented additional risk mitigations for their new models (see Box 2.2). Furthermore, AI ‘co-scientists’ are increasingly capable: they can meaningfully support top human scientists and rediscover novel, unpublished scientific findings. 575 576 Multiple research groups have developed specialised scientific AI agents capable of performing tasks including

literature review, hypothesis generation, experimental design, and data analysis. 564 574 575 576 577 578
Controlled studies and new benchmarks 33 197 549 suggest that AI systems can provide substantially more weapons development assistance than the internet alone, but larger studies are needed to confirm these results.

Evidence gaps

The primary evidence gaps relate to translating demonstrated capabilities into risk estimates. Comprehensive studies measuring how AI systems affect actual weapons development are rare, expensive, and constrained by legal and ethical considerations (see Box 2.3). Chemical risk evaluations have received relatively less attention than biological risk evaluations. 33 547 548 Across both chemical and biological risk evaluations, results are reported with varying levels of detail 579 or withheld entirely due to sensitivity concerns. Evaluations also generally assess the capabilities of individual tools, making them less applicable to real-world end-to-end workflows which might involve multiple AI systems. As such, it is unclear whether these evaluations under- or overestimate risk. Finally, there is ongoing debate about whether harmful AI capabilities primarily empower malicious actors with existing expertise (increasing their efficiency) or enable novices with little prior knowledge. 580

Mitigations

A range of technical mitigations are being developed, both within and outside of AI models, to address these risks. For general-purpose AI systems, major developers have implemented safety controls designed to refuse harmful requests. 55 581 582 Technical mitigations for specialised biological and chemical AI tools tend to lag behind those for general-purpose AI systems. 551 Other safeguards include excluding pathogen data from training, 30 55 583 584 restricting access to high-risk tools, 560 585 training models to refuse queries involving pathogenic viruses, 586 and watermarking outputs. 587 However, many of these safeguards have not been thoroughly tested, 588 and can be removed from open-weight models. 589 590 591

Another focus for technical mitigations is screening DNA synthesis requests in order to prevent malicious actors from acquiring material necessary for bioweapons creation. 570 592 593 Using synthetic DNA is likely the most straightforward way to create modified pathogens and it allows malicious actors to avoid using infectious source material. Screening is complemented by extending infectious disease surveillance frameworks to better detect novel threats and intentional attacks. 594 595 596 Biological risks – whether AI-enabled or not – can probably be at least partially mitigated by improving biosecurity directly through reducing indoor pathogen transmission, 597 developing broad-spectrum antivirals, 598 and improving laboratory biosecurity and biosafety globally. 599 600 Greater facilitation for data-sharing between relevant actors could aid in identifying and addressing potential threats. Using AI to improve pathogen detection and vaccine and drug development is likely a key mitigation strategy, especially given the limitations of current safeguards.

Challenges for policymakers

The dual-use nature of AI for biological and chemical capabilities poses challenges to policymakers wanting to limit the risk of potentially harmful uses while enabling beneficial research. The open availability of biological AI tools presents a difficult choice: whether to restrict these tools or to actively support their development for beneficial purposes 601 (see §3.4. Open-weight models).

Close

2.2. Risks from malfunctions

2.2.1. Reliability challenges

Key information

When general-purpose AI systems fail, they can cause harm. Failures include producing false or fabricated information (often described as ‘hallucinations’), writing flawed computer code, and giving misleading medical advice. These failures have the potential to cause physical or psychological harm and expose users and organisations to reputational damage, financial loss, or legal liability.

Models’ behaviour is often difficult to understand or predict, making it challenging to guarantee reliability. Even the developers of general-purpose AI models can often not meaningfully explain model behaviour, anticipate specific failure modes, or demonstrate that such failures will not occur. Malicious actors can also induce failures by interfering with AI development or giving systems adversarial inputs that evade safeguards.

AI agents pose heightened reliability risks because they act autonomously and can directly affect other systems or the physical world. Agent failures can cause greater harm because humans have fewer chances to intervene. Multi-agent systems introduce further risks, as errors can propagate and amplify through agent interactions.

Since the publication of the previous Report (January 2025), AI systems have generally become more reliable and, as a result, have seen greatly increased commercial deployment. Many kinds of failures, such as hallucinations, have generally become less likely, but systems still commonly make mistakes when performing more complex tasks.

Despite significant research efforts, no combination of methods ensures the high reliability required in critical domains. New training methods and giving AI systems access to tools can make failures less likely, but usually do not eliminate them completely.

General-purpose AI systems fail in ways that have already caused real-world harm, from fabricated legal citations to medical misdiagnoses. While human professionals also make mistakes, AI failures raise distinct concerns because of their novelty, potential scale, the difficulty of predicting when they will occur, and users' tendency to uncritically trust confident-sounding outputs. Current general-purpose AI failures include providing false information, 602 603 making basic reasoning errors, 604 605 and degrading when deployed in new contexts. 606 607 608 Documented harms from such failures include medical misdiagnoses, mistakes in legal briefs, and financial losses. 609 610 611 Reliability challenges are particularly critical for AI agents, since failures can directly cause harm without human action or oversight. 612 613 614 615 Multi-agent systems introduce further failure modes through miscoordination, conflicts, or undesired collusion between agents. 614 616

Reliability issue Examples

Hallucination

Citing non-existent precedent in legal briefs 617

Citing non-existent reduced fare policies for bereaved passengers 618

Providing inaccurate and biased medical information 619

Providing outdated information about events 620

Basic reasoning failure

Failing to perform mathematical calculations 621

Failing to infer basic causal relationships 622

Out-of-distribution failure (failure on unfamiliar or unusual inputs)

Misclassifying images when background lighting or context shifts 623

Tool use failure

Privacy breach by exposing a user's private image via an AI agent that sends it to a third-party tool 624

Failure of short-term working memory 625 626

Multi-agent system failure: miscoordination and conflict

Failing to manage shared resources because of a conflict between individual incentives and collective welfare goals 627

Table 2.4: Documented reliability issues in general-purpose AI systems, AI agents, and multi-agent systems.

General-purpose AI systems face a range of reliability challenges

Table 2.4. summarises common categories of reliability issues. The first three apply to all AI systems, while the last two pertain specifically to AI agents and multi-agent systems. Many reliability risks stem from the difficulty of predicting and monitoring AI system behaviour. These challenges (discussed further in §3.1. Technical and institutional challenges) are particularly acute for AI agents operating in complex environments. Current techniques for evaluating and mitigating such failures can reduce failure rates, but even leading AI agents are still

sufficiently unreliable to pose risks and hamper deployment in many contexts.

‘Reliability’ refers to the extent to which an AI system functions as intended by the developer or user. General-purpose AI systems experience a range of reliability issues, ranging from inaccurate or misleading content generation to failures performing basic reasoning. For example, while models have improved at recalling factual information, even leading models continue to give confident but incorrect answers at significant rates (Figure 2.10). In software engineering, general-purpose AI can now provide substantial assistance in writing, evaluating, and debugging computer code. 215 628 629 However, AI-generated code often includes bugs, 630 while coding agents regularly make errors . 631 Such failures can introduce vulnerabilities into programs and security systems (see §2.1.3. Cyberattacks).

Figure 2.10: Results of major models on the SimpleQA Verified benchmark by model release date. This benchmark measures model factuality, the ability of a model to reliably recall facts. It has a short-form question-answer (QA) format, designed to detect reliability issues such as hallucinations. Source: SimpleQA Kaggle Leaderboard, November 2025. 632

Reliability issues are particularly important to track in high-stakes settings, such as medicine, due to the accelerating use of AI and the potential for failures to result in severe harm. 609 619 Relevant capabilities have improved quickly, with leading models now able to pass medical exams. 633 634 Yet, real-world use reveals limitations that benchmarks miss. For example, in one study, models provided potentially harmful answers to 19% of medical questions posed. 635 Such failures could result in misdiagnosis, inappropriate treatment, or wrongful denial of care. 611

AI agents pose novel reliability risks due to their autonomy

Because AI agents directly act in the real world, their failures have the potential to cause more harm than failures in non-agentic systems. 99 Unlike AI systems that simply produce text or images for humans to review, AI agents can independently take actions that affect the world 99 615 636 637 (see also §1.1. What is general-purpose AI?). AI agents can initiate actions, influence other humans or AI systems, and dynamically shape future outcomes. This expanded scope of influence introduces new risks and amplifies the importance of reliability, as failures could directly cause harm with no opportunity for human intervention. 99 612 638 639 640 This may be especially important for agents deployed in strategic or safety-critical settings such as financial services, 641 energy management, 642 or scientific research. 643 644

Multi-agent AI systems introduce new kinds of reliability failures

Multi-agent AI systems introduce new kinds of reliability failures due to coordination failures or conflict between agents. In multi-agent AI systems, agents interact with each other while pursuing either shared or individual goals. 614 645 646 647 648 649 For example, in a multi-agent system designed to conduct a research literature review, a lead agent decomposes the user’s query and assigns subtasks to specialised subagents, each responsible for researching a different aspect in parallel. 650 While this allows for efficiency gains, it also means that errors can propagate between agents. 614 651 652 653 654 655 If multiple agents are built on the same base model or incorporate the same tools, then they may also exhibit correlated failures. 656 Empirical evidence for such failures in deployed systems remains limited, but these risks may grow as multi-agent systems become more common.

Figure 2.11: 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. 92

Close

Box 2.4: Deliberate attacks can also cause AI systems to fail

This section focuses on unintended reliability failures, but malicious actors can also deliberately induce failures through attacks such as prompt injections. In a prompt injection attack, malicious instructions are presented to an agent indirectly via avenues like hidden instructions in websites or databases. 507 657 658 These instructions can ‘hijack’ the agent, causing it to act against the user’s intentions. Such attacks are particularly difficult to defend against because they are delivered using external content outside the user’s or developer’s control. AI systems as targets of attack are discussed further in §2.1.3. Cyberattacks , and technical defences are covered in §3.3. Technical safeguards and monitoring .

Close

Updates

Since the publication of the last Report (January 2025), commercial and research interest in AI agents has greatly increased. More AI agents are being deployed in the real world (Figure 2.11), most of which specialise in computer-use or software engineering applications. 92 Recent releases such as XBOW hacking agent, 467 Claude-4, 659 and ChatGPT Agent 660 demonstrate nascent autonomous capabilities such as creating slide decks based on Web searches. 660 However, they cannot yet perform more complex tasks such as planning and booking travel 100 since failure rates increase for longer tasks. 98 148 Current research includes efforts to develop standards for how agents communicate with external tools and other agents. 661 662 Examples include Google's Agent2Agent 663 and Agent Payments 664 protocols, and Anthropic's Model Context Protocol. 665

Evidence gaps

The main evidence gaps stem from the difficulty of reliably evaluating AI system capabilities, limitations, and failure modes (see §3.1. Technical and institutional challenges). Systematic evaluations of the reliability of AI agents are limited and lack standardisation. 92 666 Certain issues, such as reliance on outdated information, 620 may only manifest in real-world usage, making pre-deployment evaluations inadequate. Prior work has examined the reliability of agents and multi-agent systems in conventional software and earlier forms of AI. 647 667 668 However, the applicability of this work to modern AI agents, which are often based on large language models, is unclear. 669 Some researchers have raised concerns about the novel behaviours agents may exhibit in their interactions with each other, such as collusion or correlated failures, 614 but empirical evidence remains limited. Efforts to address these gaps include the National Institute of Standards and Technology's (NIST's) new evaluations of agent-hijacking risks, 670 the OECD's AI Capability Indicators (243), and UK AI Security Institute's Inspect Sandboxing Toolkit. 671

Mitigations

Techniques for improving AI reliability target both the model itself and the broader system in which it is deployed. These can reduce failure rates, but none can yet ensure the high reliability required in critical domains. 672 An important technical measure is adversarial training, which exposes models to challenging inputs during training to help it develop more suitable, robust responses 673 674 675 676 677 (see §3.3. Technical safeguards and monitoring). To reduce hallucinations, developers can apply retrieval-augmented generation (RAG), which supplements a model's responses with information retrieved from an external database, helping ensure outputs are accurate and current, 678 679 680 or specifically fine-tune models to be more factual 681 or reason more effectively. 682 Environment- or tool-based methods can also help developers monitor AI systems. 683 For example, deployers could pilot AI systems in limited sandboxed environments to analyse potential failure modes before deploying them more broadly.

For AI agents specifically, researchers have proposed improving reliability through improved transparency, oversight, and monitoring. For example, monitoring agents' interactions with external tools and with other agents would allow for more effective oversight of agent activities 684 685 and incident analysis. 686 Methods for collecting such information automatically, including in multi-agent settings, remain an active area of research. 653 654

Challenges for policymakers

Key challenges for policymakers include weighing the benefits of AI agent deployment against the risks of reliability failures, and ensuring that developers, deployers, and users have access to accurate information about agent performance and risk profiles. Deciding how to attribute liability for harms caused by AI agents poses a further challenge, 639 particularly in multi-agent settings where it may be hard to identify when and how failures occurred. 687 These challenges are compounded by the difficulty of evaluating agent reliability as agents gain autonomy and access to external tools. 688 689 Uncertainty about how quickly agentic capabilities will emerge also makes planning for novel challenges difficult (see §3.1. Technical and institutional challenges regarding the 'evidence dilemma').

Close

2.2.2. Loss of control

Key information

Loss of control scenarios are scenarios in which one or more general-purpose AI systems operate outside of anyone's control, and regaining control is either extremely costly or impossible. These hypothesised scenarios vary in their severity, but some experts give credence to outcomes as severe as the marginalisation or extinction of humanity.

Expert opinion on the likelihood of loss of control varies greatly. Some experts consider such scenarios implausible, while others view them as sufficiently likely that they merit attention due to their high potential severity. Disagreement about this risk overall stems from disagreements about future AI capabilities, behavioural propensities, and deployment trajectories.

Current AI systems show early signs of relevant capabilities, but not at levels that would enable loss of control. Systems would need a range of advanced capabilities to cause loss of control, including the ability to evade oversight, execute long-term plans, and prevent deployers and other actors from implementing countermeasures.

Loss of control becomes more likely if AI systems are 'misaligned', meaning they have goals that conflict with the intentions of developers, users, or society more broadly. To continue pursuing such goals, a misaligned system might provide false information, conceal undesirable actions, or resist shutdown.

Since the publication of the previous Report (January 2025), models have shown more advanced planning and oversight-undermining capabilities, making it more difficult to evaluate their capabilities. Models have improved at 'reward hacking' their evaluations by finding loopholes and now regularly identify evaluation prompts as tests, a capability known as 'situational awareness'.

Managing potential loss of control could require substantial advance preparation despite existing uncertainties. A key challenge for policymakers is preparing for a risk whose likelihood, nature, and timing remains unusually ambiguous.

Loss of control scenarios involve one or more general-purpose AI systems coming to operate outside of anyone's control, with regaining control being either extremely costly or impossible. Concerns about loss of control have deep historical roots, 690 691 692 693 694 having been raised by foundational figures in computing such as Alan Turing, I. J. Good, and Norbert Wiener. 695 696 697 Recent improvements in capabilities (see §1.2. Current capabilities) have revived them. 698 699 700 This section examines three factors that would need to be present for such scenarios to occur: whether AI systems will develop capabilities that could significantly undermine human control; whether they develop a propensity to use such capabilities harmfully; and whether they are deployed in environments that provide opportunities to do so.

Experts disagree about the likelihood and potential severity of loss of control scenarios. 701 702 Some believe that outcomes as extreme as the extinction of humanity are plausible. 700 703 704 705 706 707 Others think that such catastrophic outcomes are implausible, arguing that AI systems will never develop the necessary capabilities or that monitoring mechanisms will identify and prevent dangerous behaviours. 708 709 710 711 Loss of control can therefore be understood as a risk with uncertain likelihood but potentially extreme severity.

Hypothesised loss of control scenarios vary in how severe and widespread their effects are and how quickly they manifest. 102 698 700 712 713 714 This section focuses on particularly severe scenarios where regaining control would be extremely costly or impossible. These are different from current instances of AI behaving in unintended or undesirable ways (see §2.2.1. Reliability challenges) . * Present-day AI systems sometimes produce outputs that conflict with developer or user intentions. By contrast, the loss of control scenarios discussed here would require AI systems to not only possess substantially greater capabilities, but also to deploy those capabilities in sophisticated ways to undermine oversight measures. Three factors that would allow such scenarios to occur:

Sufficient capabilities: AI systems must develop capabilities that could allow them to undermine human control.

Harmful propensity: AI systems must exhibit a propensity to actually leverage these capabilities in ways that lead to a loss of control.

Enabling deployment environment: Humans must deploy such systems in contexts where they have or can gain the access and opportunity to cause harm.

The remainder of this section discusses these factors, as well as the effectiveness of oversight mechanisms to identify and control AI systems that may pose a risk of loss of control.

* This section focuses on active loss of control scenarios. 50 This is distinct from passive loss of control scenarios, where the broad adoption of AI systems undermines human control through over-reliance on AI for decision-making or other important societal functions (similar scenarios are partly discussed in §2.3.2. Risks to human autonomy).

What capabilities could enable loss of control scenarios?

AI systems would need to possess a range of advance