AI ACTION SUMMIT

International AI Safety Report

The International Scientific Report on the Safety of Advanced AI

January 2025

Contributors

CHAIR

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

EXPERT ADVISORY PANEL

This international panel was nominated by the governments of the 30 countries listed below, the UN, EU, and OECD.

Australia: Bronwyn Fox, the University of New South Wales

Brazil: André Carlos Ponce de Leon Ferreira de Carvalho, Institute of Mathematics and Computer Sciences, University of São Paulo

Canada: Mona Nemer, Chief Science Advisor of Canada

Chile: Raquel Pezoa Rivera, Universidad Técnica Federico Santa Maria

China: Yi Zeng, Chinese Academy of Sciences

European Union: Juha Heikkilä, European AI Office

France: Guillaume Avrin, National Coordination for Artificial Intelligence

Germany: Antonio Krüger, German Research Center for Artificial Intelligence

India: Balaraman Ravindran, Wadhwani School of Data Science and AI, Indian Institute of Technology Madras

Indonesia: Hammam Riza, Collaborative Research and Industrial Innovation in Artificial Intelligence (KORIKA)

Ireland: Ciarán Seoighe, Research Ireland

Israel: Ziv Katzir, Israel Innovation Authority

Italy: Andrea Monti, Legal Expert for the Undersecretary of State for the Digital Transformation, Italian Ministers Council's Presidency

Japan: Hiroaki Kitano, Sony Group Corporation

Kenya: Nusu Mwamanzi, Ministry of ICT & Digital Economy

Kingdom of Saudi Arabia: Fahad Albalawi, Saudi Authority for Data and Artificial Intelligence

Mexico: José Ramón López Portillo, LobsterTel

Netherlands: Haroon Sheikh, Netherlands' Scientific Council for Government Policy

New Zealand: Gill Jolly, Ministry of Business, Innovation and Employment

Nigeria: Olubunmi Ajala, Ministry of Communications, Innovation and Digital Economy

OECD: Jerry Sheehan, Director of the Directorate for Science, Technology and Innovation

Philippines: Dominic Vincent Ligot, CirroLytix

Republic of Korea: Kyoung Mu Lee, Department of Electrical and Computer Engineering, Seoul National University

Rwanda: Crystal Rugege, Centre for the Fourth Industrial Revolution

Singapore: Denise Wong, Data Innovation and Protection Group, Infocomm Media Development Authority

Spain: Nuria Oliver, ELLIS Alicante

Switzerland: Christian Busch, Federal Department of Economic Affairs, Education and Research

Türkiye: Ahmet Halit Hatip, Turkish Ministry of Industry and Technology

Ukraine: Oleksii Molchanovskyi, Expert Committee on the Development of Artificial Intelligence in Ukraine

United Arab Emirates: Marwan Alserkal, Ministry of Cabinet Affairs, Prime Minister's Office

United Kingdom: Chris Johnson, Chief Scientific Adviser in the Department for Science, Innovation and Technology

United Nations: Amandeep Singh Gill, Under-Secretary-General for Digital and Emerging Technologies and Secretary-General's Envoy on Technology

United States: Saif M. Khan, U.S. Department of Commerce

SCIENTIFIC LEAD

Sören Mindermann, Mila - Quebec AI Institute

LEAD WRITER

Daniel Privitera, KIRA Center

WRITING GROUP

Tamay Besiroglu, Epoch AI

Rishi Bommasani, Stanford University

Stephen Casper, Massachusetts Institute of Technology

Yejin Choi, Stanford University

Philip Fox, KIRA Center

Ben Garfinkel, University of Oxford

Danielle Goldfarb, Mila - Quebec AI Institute

Hoda Heidari, Carnegie Mellon University

Anson Ho, Epoch AI

Sayash Kapoor, Princeton University

Leila Khalatbari, Hong Kong University of Science and Technology

Shayne Longpre, Massachusetts Institute of Technology

Sam Manning, Centre for the Governance of AI

Vasilios Mavroudis, The Alan Turing Institute

Mantas Mazeika, University of Illinois at Urbana-Champaign

Julian Michael, New York University

Jessica Newman, University of California, Berkeley

Kwan Yee Ng, Concordia AI

Chinasa T. Okolo, Brookings Institution

Deborah Raji, University of California, Berkeley

Girish Sastry, Independent

Elizabeth Seger (generalist writer), Demos

Theodora Skeadas, Humane Intelligence

Tobin South, Massachusetts Institute of Technology

Emma Strubell, Carnegie Mellon University

Florian Tramèr, ETH Zurich

Lucia Velasco, Maastricht University

Nicole Wheeler, University of Birmingham

SENIOR ADVISERS

Daron Acemoglu, Massachusetts Institute of Technology

Olubayo Adekanmbi, contributed as a Senior Adviser prior to taking up his role at EqualyzAI

David Dalrymple, Advanced Research + Invention Agency

Thomas G. Dietterich, Oregon State University

Edward W. Felten, Princeton University

Pascale Fung, contributed as a Senior Adviser prior to taking up her role at Meta

Pierre-Olivier Gourinchas, Research Department, International Monetary Fund

Fredrik Heintz, Linköping University

Geoffrey Hinton, University of Toronto

Nick Jennings, University of Loughborough

Andreas Krause, ETH Zurich

Susan Leavy, University College Dublin

Percy Liang, Stanford University

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

Alice Oh, KAIST School of Computing

Gopal Ramchurn, Responsible AI UK

Stuart Russell, University of California, Berkeley

Marietje Schaake, Stanford University

Bernhard Schölkopf, ELLIS Institute Tübingen

Da**wn Song,** University of California, Berkeley

Alvaro Soto, Pontificia Universidad Católica de Chile

Lee Tiedrich, Duke University

Gaël Varoquaux, Inria

Andrew Yao, Institute for Interdisciplinary Information Sciences, Tsinghua University

Ya-Qin Zhang, Tsinghua University

SECRETARIAT

AI Safety Institute

Baran Acar

Ben Clifford

Lambrini Das

Claire Dennis

Freya Hempleman

Hannah Merchant

Rian Overy

Ben Snodin

Mila — Quebec AI Institute

Jonathan Barry

Benjamin Prud'homme

ACKNOWLEDGEMENTS

Civil Society and Industry Reviewers

Civil Society: Ada Lovelace Institute, AI Forum New Zealand / Te Kāhui Atamai Iahiko o Aotearoa, Australia's Temporary AI Expert Group, 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, Eon Institute, Gradient Institute, Israel Democracy Institute, Mozilla Foundation, Old Ways New, RAND, SaferAI, The Centre for Long-Term Resilience, The Future Society, The Alan Turing Institute, The Royal Society, Türkiye Artificial Intelligence Policies Association.

Industry: Advai, Anthropic, Cohere, Deloitte Consulting USA and Deloitte LLM UK, G42, Google DeepMind, Harmony Intelligence, Hugging Face, IBM, Lelapa AI, Meta, Microsoft, Shutterstock, Zhipu.ai.

Special Thanks

The Secretariat appreciates the support, comments and feedback from Angie Abdilla, Concordia AI, Nitarshan Rajkumar, Geoffrey Irving, Shannon Vallor, Rebecca Finlay and Andrew Strait.

© Crown owned 2025

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: psi@nationalarchives.gsi.gov.uk.

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: secretariat.AIStateofScience@dsit.gov.uk.

Enquiries regarding the content of the report should also be sent to the Scientific Lead.

Disclaimer

The report does not 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. This report is a synthesis of the existing research on the capabilities and risks of advanced AI. The Chair of the report has ultimate responsibility for it and has overseen its development from beginning to end.

Research series number: DSIT 2025/001

About this report

- **This is the first International AI Safety Report.** Following an interim publication in May 2024, a diverse group of 96 Artificial Intelligence (AI) experts contributed to this first full report, including an international Expert Advisory Panel nominated by 30 countries, the Organisation for Economic Co-operation and Development (OECD), the European Union (EU), and the United Nations (UN). The report aims to provide scientific information that will support informed policymaking. It does not recommend specific policies.
- **The report is the work of independent experts.** Led by the Chair, the independent experts writing this report collectively had full discretion over its content.
- **While this report is concerned with AI risks and AI safety, AI also offers many potential benefits for people, businesses, and society.** There are many types of AI, each with different benefits and risks. Most of the time, in most applications, AI helps individuals and organisations be more effective. But people around the world will only be able to fully enjoy AI's many potential benefits safely if its risks are appropriately managed. This report focuses on identifying these risks and evaluating methods for mitigating them. It does not aim to comprehensively assess all possible societal impacts of AI, including its many potential benefits.
- **The focus of the report is general-purpose AI.** The report restricts its focus to a type of AI that has advanced particularly rapidly in recent years, and whose associated risks have been less studied and understood: general-purpose AI, or AI that can perform a wide variety of tasks. The analysis in this report focuses on the most advanced general-purpose AI systems at the time of writing, as well as future systems that might be even more capable.
- **The report summarises the scientific evidence on three core questions:** What can general-purpose AI do? What are risks associated with general-purpose AI? And what mitigation techniques are there against these risks?
- **The stakes are high.** We, the experts contributing to this report, continue to disagree on several questions, minor and major, around general-purpose AI capabilities, risks, and risk mitigations. But we consider this report essential for improving our collective understanding of this technology and its potential risks. We hope that the report will help the international community to move towards greater consensus about general-purpose AI and mitigate its risks more effectively, so that people can safely experience its many potential benefits. The stakes are high. We look forward to continuing this effort.

Update on latest AI advances after the writing of this report: Chair's note

Between the end of the writing period for this report (5 December 2024) and the publication of this report in January 2025, an important development took place. The AI company OpenAI shared early test results from a new AI model, *o3*. These results indicate significantly stronger performance than any previous model on a number of the field's most challenging tests of programming, abstract reasoning, and scientific reasoning. In some of these tests, o3 outperforms many (but not all) human experts. Additionally, it achieves a breakthrough on a key abstract reasoning test that many experts, including myself, thought was out of reach until recently. However, at the time of writing there is no public information about its real-world capabilities, particularly for solving more open-ended tasks.

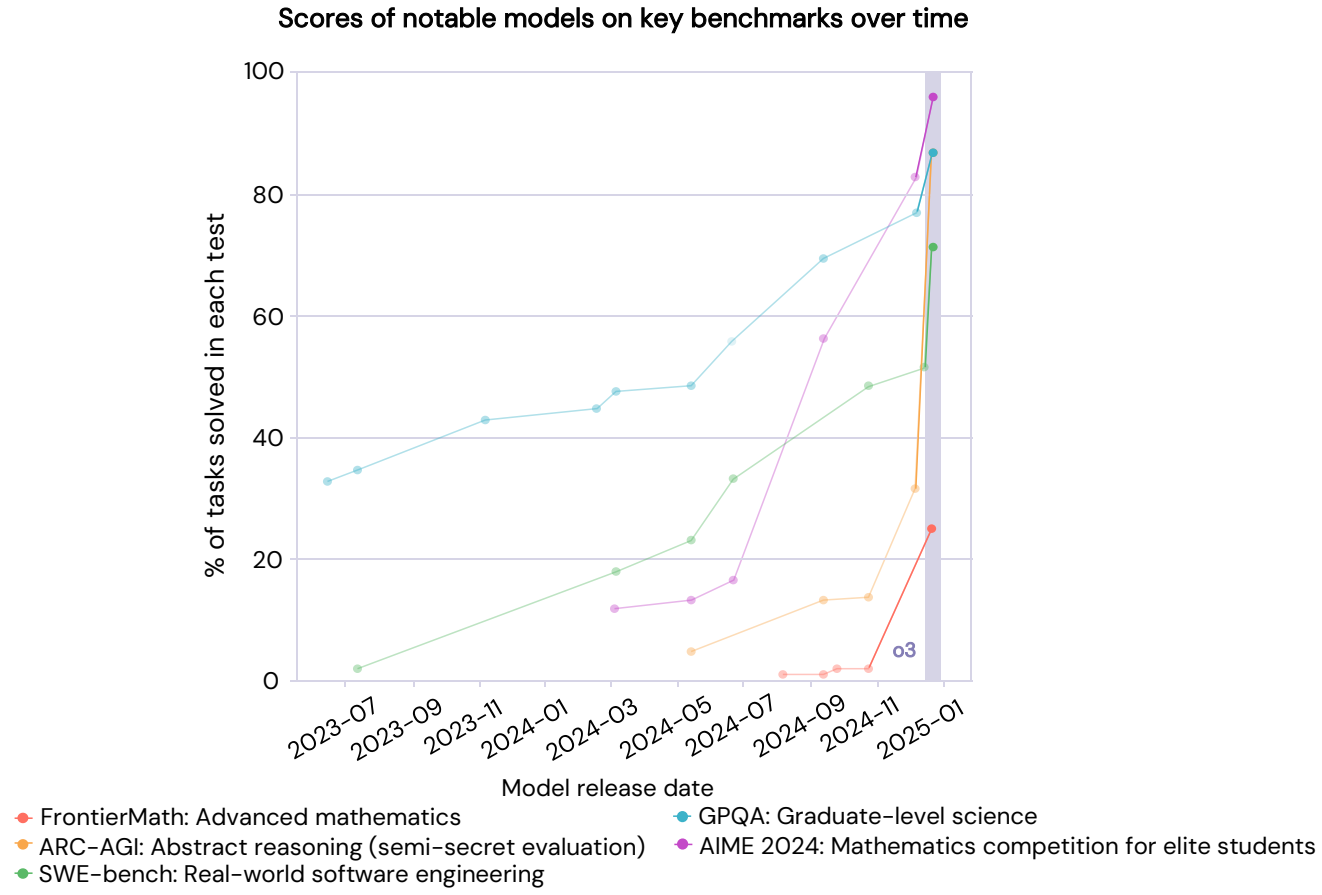


Figure 0.1: *Scores of notable general-purpose AI models on key benchmarks from June 2023 to December 2024. o3 showed significantly improved performance compared to the previous state of the art (shaded region). These benchmarks are some of the field's most challenging tests of programming, abstract reasoning, and scientific reasoning. For the unreleased o3, the announcement date is shown; for the other models, the release date is shown. Some of the more recent AI models, including o3, benefited from improved scaffolding and more computation at test-time. Sources: Anthropic, 2024; Chollet, 2024; Chollet et al., 2025; Epoch AI, 2024; Glazer et al. 2024; OpenAI, 2024a; OpenAI, 2024b; Jimenez et al., 2024; Jimenez et al., 2025.*

The o3 results are evidence that the pace of advances in AI capabilities may remain high or even accelerate. More specifically, they suggest that giving models more computing power for solving a given problem ('inference scaling') may help overcome previous limitations. Generally speaking, inference scaling makes models more expensive to use. But as another recent notable model, *R1*, released by the company DeepSeek in January 2025, has shown, researchers are successfully working on lowering these costs. Overall, inference scaling may allow AI developers to make further advances going forward. The o3 results also underscore the need to better understand how AI developers' growing use of AI may affect the speed of further AI development itself.

The trends evidenced by o3 could have profound implications for AI risks. Advances in science and programming capabilities have previously generated more evidence for risks such as cyber and biological attacks. The o3 results are also relevant to potential labour market impacts, loss of control risk, and energy use among others. But o3's capabilities could also be used to help protect against malfunctions and malicious uses. Overall, the risk assessments in this report should be read with the understanding that AI has gained capabilities since the report was written. However, so far there is no evidence yet about o3's real world impacts, and no information to confirm nor rule out major novel and/or immediate risks.

The improvement in capabilities suggested by the o3 results and our limited understanding of the implications for AI risks underscore a key challenge for policymakers that this report identifies: they will often have to weigh potential benefits and risks of imminent AI advancements without having a large body of scientific evidence available. Nonetheless, generating evidence on the safety and security implications of the trends implied by o3 will be an urgent priority for AI research in the coming weeks and months.

Key findings of the report

- **The capabilities of general-purpose AI, the type of AI that this report focuses on, have increased rapidly in recent years and have improved further in recent months.**[†] A few years ago, the best large language models (LLMs) could rarely produce a coherent paragraph of text. Today, general-purpose AI can write computer programs, generate custom photorealistic images, and engage in extended open-ended conversations. Since the publication of the Interim Report (May 2024), new models have shown markedly better performance at tests of scientific reasoning and programming.
- **Many companies are now investing in the development of general-purpose AI agents, as a potential direction for further advancement.** AI agents are general-purpose AI systems which can autonomously act, plan, and delegate to achieve goals with little to no human oversight. Sophisticated AI agents would be able to, for example, use computers to complete longer projects than current systems, unlocking both additional benefits and additional risks.
- **Further capability advancements in the coming months and years could be anything from slow to extremely rapid.**[†] Progress will depend on whether companies will be able to rapidly deploy even more data and computational power to train new models, and whether 'scaling' models in this way will overcome their current limitations. Recent research suggests that rapidly scaling up models may remain physically feasible for at least several years. But major capability advances may also require other factors: for example, new research breakthroughs, which are hard to predict, or the success of a novel scaling approach that companies have recently adopted.
- **Several harms from general-purpose AI are already well established.** These include scams, non-consensual intimate imagery (NCII) and child sexual abuse material (CSAM), model outputs that are biased against certain groups of people or certain opinions, reliability issues, and privacy violations. Researchers have developed mitigation techniques for these problems, but so far no combination of techniques can fully resolve them. Since the publication of the Interim Report, new evidence of discrimination related to general-purpose AI systems has revealed more subtle forms of bias.
- **As general-purpose AI becomes more capable, evidence of additional risks is gradually emerging.** These include risks such as large-scale labour market impacts, AI-enabled hacking or biological attacks, and society losing control over general-purpose AI. Experts interpret the existing evidence on these risks differently: some think that such risks are decades away, while others think that general-purpose AI could lead to societal-scale harm within the next few years. Recent advances in general-purpose AI capabilities – particularly in tests of scientific reasoning and programming – have generated new evidence for potential risks such as AI-enabled hacking and biological attacks, leading one major AI company to increase its assessment of biological risk from its best model from 'low' to 'medium'.

[†] Please refer to the Chair's update on the latest AI advances after the writing of this report.

- **Risk management techniques are nascent, but progress is possible.** There are various technical methods to assess and reduce risks from general-purpose AI that developers can employ and regulators can require, but they all have limitations. For example, current interpretability techniques for explaining why a general-purpose AI model produced any given output remain severely limited. However, researchers are making some progress in addressing these limitations. In addition, researchers and policymakers are increasingly trying to standardise risk management approaches, and to coordinate internationally.
- **The pace and unpredictability of advancements in general-purpose AI pose an 'evidence dilemma' for policymakers.**[†] Given sometimes rapid and unexpected advancements, policymakers will often have to weigh potential benefits and risks of imminent AI advancements without having a large body of scientific evidence available. In doing so, they face a dilemma. On the one hand, pre-emptive risk mitigation measures based on limited evidence might turn out to be ineffective or unnecessary. On the other hand, waiting for stronger evidence of impending risk could leave society unprepared or even make mitigation impossible – for instance if sudden leaps in AI capabilities, and their associated risks, occur. Companies and governments are developing early warning systems and risk management frameworks that may reduce this dilemma. Some of these trigger specific mitigation measures when there is new evidence of risks, while others require developers to provide evidence of safety before releasing a new model.
- **There is broad consensus among researchers that advances regarding the following questions would be helpful:** How rapidly will general-purpose AI capabilities advance in the coming years, and how can researchers reliably measure that progress? What are sensible risk thresholds to trigger mitigations? How can policymakers best gain access to information about general-purpose AI that is relevant to public safety? How can researchers, technology companies, and governments reliably assess the risks of general-purpose AI development and deployment? How do general-purpose AI models work internally? How can general-purpose AI be designed to behave reliably?
- **AI does not happen to us: choices made by people determine its future.** The future of general-purpose AI technology is uncertain, with a wide range of trajectories appearing to be possible even in the near future, including both very positive and very negative outcomes. This uncertainty can evoke fatalism and make AI appear as something that happens to us. But it will be the decisions of societies and governments on how to navigate this uncertainty that determine which path we will take. This report aims to facilitate constructive and evidence-based discussion about these decisions.

[†] Please refer to the Chair's update on the latest AI advances after the writing of this report.

Executive Summary

The purpose of this report

This report synthesises the state of scientific understanding of general-purpose AI – AI that can perform a wide variety of tasks – with a focus on understanding and managing its risks.

This report summarises the scientific evidence on the safety of general-purpose AI. The purpose of this report is to help create a shared international understanding of risks from advanced AI and how they can be mitigated. To achieve this, this report focuses on general-purpose AI – or AI that can perform a wide variety of tasks – since this type of AI has advanced particularly rapidly in recent years and has been deployed widely by technology companies for a range of consumer and business purposes. The report synthesises the state of scientific understanding of general-purpose AI, with a focus on understanding and managing its risks.

Amid rapid advancements, research on general-purpose AI is currently in a time of scientific discovery, and – in many cases – is not yet settled science. The report provides a snapshot of the current scientific understanding of general-purpose AI and its risks. This includes identifying areas of scientific consensus and areas where there are different views or gaps in the current scientific understanding.

People around the world will only be able to fully enjoy the potential benefits of general-purpose AI safely if its risks are appropriately managed. This report focuses on identifying those risks and evaluating technical methods for assessing and mitigating them, including ways that general-purpose AI itself can be used to mitigate risks. It does not aim to comprehensively assess all possible societal impacts of general-purpose AI. Most notably, the current and potential future benefits of general-purpose AI – although they are vast – are beyond this report's scope. Holistic policymaking requires considering both the potential benefits of general-purpose AI and the risks covered in this report. It also requires taking into account that other types of AI have different risk/benefit profiles compared to current general-purpose AI.

The three main sections of the report summarise the scientific evidence on three core questions: What can general-purpose AI do? What are risks associated with general-purpose AI? And what mitigation techniques are there against these risks?

Section 1 – Capabilities of general-purpose AI: What can general-purpose AI do now and in the future?

General-purpose AI capabilities have improved rapidly in recent years, and further advancements could be anything from slow to extremely rapid.

What AI can do is a key contributor to many of the risks it poses, and according to many metrics, general-purpose AI capabilities have been progressing rapidly. Five years ago, the leading general-purpose AI language models could rarely produce a coherent paragraph of text. Today, some general-purpose AI models can engage in conversations on a wide range of topics, write computer programs, or generate realistic short videos from a description. However, it is technically challenging to reliably estimate and describe the capabilities of general-purpose AI.

AI developers have rapidly improved the capabilities of general-purpose AI in recent years, mostly through 'scaling'.[†] They have continually increased the resources used for training new models (this is often referred to as 'scaling') and refined existing approaches to use those resources more efficiently. For example, according to recent estimates, state-of-the-art AI models have seen annual increases of approximately 4x in computational resources ('compute') used for training and 2.5x in training dataset size.

The pace of future progress in general-purpose AI capabilities has substantial implications for managing emerging risks, but experts disagree on what to expect even in the coming months and years. Experts variously support the possibility of general-purpose AI capabilities advancing slowly, rapidly, or extremely rapidly.

Experts disagree about the pace of future progress because of different views on the promise of further 'scaling' – and companies are exploring an additional, new type of scaling that might further accelerate capabilities.[†] While scaling has often overcome the limitations of previous systems, experts disagree about its potential to resolve the remaining limitations of today's systems, such as unreliability at acting in the physical world and at executing extended tasks on computers. In recent months, a new type of scaling has shown potential for further improving capabilities: rather than just scaling up the resources used for training models, AI companies are also increasingly interested in 'inference scaling' – letting an already trained model use more computation to solve a given problem, for example to improve on its own solution, or to write so-called 'chains of thought' that break down the problem into simpler steps.

Several leading companies that develop general-purpose AI are betting on 'scaling' to continue leading to performance improvements. If recent trends continue, by the end of 2026 some

[†] Please refer to the Chair's update on the latest AI advances after the writing of this report.

general-purpose AI models will be trained using roughly 100x more training compute than 2023's most compute-intensive models, growing to 10,000x more training compute by 2030, combined with algorithms that achieve greater capabilities for a given amount of available computation. In addition to this potential scaling of training resources, recent trends such as inference scaling and using models to generate training data could mean that even more compute will be used overall. However, there are potential bottlenecks to further increasing both data and compute rapidly, such as the availability of data, AI chips, capital, and local energy capacity. Companies developing general-purpose AI are working to navigate these potential bottlenecks.

Since the publication of the Interim Report (May 2024), general-purpose AI has reached expert-level performance in some tests and competitions for scientific reasoning and programming, and companies have been making large efforts to develop autonomous AI agents. Advances in science and programming have been driven by inference scaling techniques such as writing long 'chains of thought'. New studies suggest that further scaling such approaches, for instance allowing models to analyse problems by writing even longer chains of thought than today's models, could lead to further advances in domains where reasoning matters more, such as science, software engineering, and planning. In addition to this trend, companies are making large efforts to develop more advanced general-purpose AI agents, which can plan and act autonomously to work towards a given goal. Finally, the market price of using general-purpose AI of a given capability level has dropped sharply, making this technology more broadly accessible and widely used.

This report focuses primarily on technical aspects of AI progress, but how fast general-purpose AI will advance is not a purely technical question. The pace of future advancements will also depend on non-technical factors, potentially including the approaches that governments take to regulating AI. This report does not discuss how different approaches to regulation might affect the speed of development and adoption of general-purpose AI.

Section 2 – Risks: What are risks associated with general-purpose AI?

Several harms from general-purpose AI are already well-established. As general-purpose AI becomes more capable, evidence of additional risks is gradually emerging.

This report classifies general-purpose AI risks into three categories: malicious use risks; risks from malfunctions; and systemic risks. Each of these categories contains risks that have already materialised as well as risks that might materialise in the next few years.

Risks from malicious use: **malicious actors can use general-purpose AI to cause harm to individuals, organisations, or society.** Forms of malicious use include:

- **Harm to individuals through fake content:** Malicious actors can currently use general-purpose AI to generate fake content that harms individuals in a targeted way. These malicious uses include non-consensual 'deepfake' pornography and AI-generated CSAM, financial fraud through voice impersonation, blackmail for extortion, sabotage of personal and professional reputations, and psychological abuse. However, while incident reports of harm from AI-generated fake content are common, reliable statistics on the frequency of these incidents are still lacking.
- **Manipulation of public opinion:** General-purpose AI makes it easier to generate persuasive content at scale. This can help actors who seek to manipulate public opinion, for instance to affect political outcomes. However, evidence on how prevalent and how effective such efforts are remains limited. Technical countermeasures like content watermarking, although useful, can usually be circumvented by moderately sophisticated actors.
- **Cyber offence:** General-purpose AI can make it easier or faster for malicious actors of varying skill levels to conduct cyberattacks. Current systems have demonstrated capabilities in low- and medium-complexity cybersecurity tasks, and state-sponsored actors are actively exploring AI to survey target systems. New research has confirmed that the capabilities of general-purpose AI related to cyber offence are significantly advancing, but it remains unclear whether this will affect the balance between attackers and defenders.
- **Biological and chemical attacks:** Recent general-purpose AI systems have displayed some ability to provide instructions and troubleshooting guidance for reproducing known biological and chemical weapons and to facilitate the design of novel toxic compounds. In new experiments that tested for the ability to generate plans for producing biological weapons, a general-purpose AI system sometimes performed better than human experts with access to the internet. In response, one AI company increased its assessment of biological risk from its best model from 'low' to 'medium'. However, real-world attempts to develop such weapons still require substantial additional resources and expertise. A comprehensive assessment of biological and chemical risk is difficult because much of the relevant research is classified.

Since the publication of the Interim Report, general-purpose AI has become more capable in domains that are relevant for malicious use. For example, researchers have recently built general-purpose AI systems that were able to find and exploit some cybersecurity vulnerabilities on their own and, with human assistance, discover a previously unknown vulnerability in widely used software. General-purpose AI capabilities related to reasoning and to integrating different types of data, which can aid research on pathogens or in other dual-use fields, have also improved.

Risks from malfunctions: **general-purpose AI can also cause unintended harm.** Even when users have no intention to cause harm, serious risks can arise due to the malfunctioning of general-purpose AI. Such malfunctions include:

- **Reliability issues:** Current general-purpose AI can be unreliable, which can lead to harm. For example, if users consult a general-purpose AI system for medical or legal advice, the system might generate an answer that contains falsehoods. Users are often not aware of the limitations of an AI product, for example due to limited 'AI literacy', misleading advertising, or miscommunication. There are a number of known cases of harm from reliability issues, but still limited evidence on exactly how widespread different forms of this problem are.
- **Bias:** General-purpose AI systems can amplify social and political biases, causing concrete harm. They frequently display biases with respect to race, gender, culture, age, disability, political opinion, or other aspects of human identity. This can lead to discriminatory outcomes including unequal resource allocation, reinforcement of stereotypes, and systematic neglect of underrepresented groups or viewpoints. Technical approaches for mitigating bias and discrimination in general-purpose AI systems are advancing, but face trade-offs between bias mitigation and competing objectives such as accuracy and privacy, as well as other challenges.
- **Loss of control:** 'Loss of control' scenarios are hypothetical future scenarios in which one or more general-purpose AI systems come to operate outside of anyone's control, with no clear path to regaining control. There is broad consensus that current general-purpose AI lacks the capabilities to pose this risk. However, expert opinion on the likelihood of loss of control within the next several years varies greatly: some consider it implausible, some consider it likely to occur, and some see it as a modest-likelihood risk that warrants attention due to its high potential severity. Ongoing empirical and mathematical research is gradually advancing these debates.

Since the publication of the Interim Report, new research has led to some new insights about risks of bias and loss of control. The evidence of bias in general-purpose AI systems has increased, and recent work has detected additional forms of AI bias. Researchers have observed modest further advancements towards AI capabilities that are likely necessary for commonly discussed loss of control scenarios to occur. These include capabilities for autonomously using computers, programming, gaining unauthorised access to digital systems, and identifying ways to evade human oversight.

***Systemic risks:* beyond the risks directly posed by capabilities of individual models, widespread deployment of general-purpose AI is associated with several broader systemic risks.** Examples of systemic risks range from potential labour market impacts to privacy risks and environmental effects:

- **Labour market risks:** General-purpose AI, especially if it continues to advance rapidly, has the potential to automate a very wide range of tasks, which could have a significant effect on the labour market. This means that many people could lose their current jobs. However, many economists expect that potential job losses could be offset, partly or potentially even completely, by the creation of new jobs and by increased demand in non-automated sectors.

- **Global AI R&D divide:** General-purpose AI research and development (R&D) is currently concentrated in a few Western countries and China. This 'AI divide' has the potential to increase much of the world's dependence on this small set of countries. Some experts also expect it to contribute to global inequality. The divide has many causes, including a number of causes that are not unique to AI. However, in significant part it stems from differing levels of access to the very expensive compute needed to develop general-purpose AI: most low- and middle-income countries (LMICs) have significantly less access to compute than high-income countries (HICs).
- **Market concentration and single points of failure:** A small number of companies currently dominate the market for general-purpose AI. This market concentration could make societies more vulnerable to several systemic risks. For instance, if organisations across critical sectors, such as finance or healthcare, all rely on a small number of general-purpose AI systems, then a bug or vulnerability in such a system could cause simultaneous failures and disruptions on a broad scale.
- **Environmental risks:** Growing compute use in general-purpose AI development and deployment has rapidly increased the amounts of energy, water, and raw material consumed in building and operating the necessary compute infrastructure. This trend shows no clear indication of slowing, despite progress in techniques that allow compute to be used more efficiently. General-purpose AI also has a number of applications that can either benefit or harm sustainability efforts.
- **Privacy risks:** General-purpose AI can cause or contribute to violations of user privacy. For example, sensitive information that was in the training data can leak unintentionally when a user interacts with the system. In addition, when users share sensitive information with the system, this information can also leak. But general-purpose AI can also facilitate deliberate violations of privacy, for example if malicious actors use AI to infer sensitive information about specific individuals from large amounts of data. However, so far, researchers have not found evidence of widespread privacy violations associated with general-purpose AI.
- **Copyright infringements:** General-purpose AI both learns from and creates works of creative expression, challenging traditional systems of data consent, compensation, and control. Data collection and content generation can implicate a variety of data rights laws, which vary across jurisdictions and may be under active litigation. Given the legal uncertainty around data collection practices, AI companies are sharing less information about the data they use. This opacity makes third-party AI safety research harder.

Since the publication of the Interim Report, additional evidence on the labour market impacts of general-purpose AI has emerged, while new developments have heightened privacy and copyrights concerns. New analyses of labour market data suggest that individuals are adopting general-purpose AI very rapidly relative to previous technologies. The pace of adoption by businesses varies widely by sector. In addition, recent advances in capabilities have led to general-purpose AI being deployed increasingly in sensitive contexts such as healthcare or workplace monitoring, which creates new privacy risks. Finally, as copyright disputes intensify

and technical mitigations to copyright infringements remain unreliable, data rights holders have been rapidly restricting access to their data.

***Open-weight models:* an important factor in evaluating many risks that a general-purpose AI model might pose is how it is released to the public.** So-called 'open-weight models' are AI models whose central components, called 'weights', are shared publicly for download. Open-weight access facilitates research and innovation, including in AI safety, as well as increasing transparency and making it easier for the research community to detect flaws in models. However, open-weight models can also pose risks, for example by facilitating malicious or misguided use that is difficult or impossible for the developer of the model to monitor or mitigate. Once model weights are available for public download, there is no way to implement a wholesale rollback of all existing copies or ensure that all existing copies receive safety updates. Since the Interim Report, high-level consensus has emerged that risks posed by greater AI openness should be evaluated in terms of 'marginal' risk: the extent to which releasing an open-weight model would increase or decrease a given risk, relative to risks posed by existing alternatives such as closed models or other technologies.

Section 3 – Risk management: What techniques are there for managing risks from general-purpose AI?

Several technical approaches can help manage risks, but in many cases the best available approaches still have highly significant limitations and no quantitative risk estimation or guarantees that are available in other safety-critical domains.

Risk management – identifying and assessing risks, and then mitigating and monitoring them – is difficult in the context of general-purpose AI. Although risk management has also been highly challenging in many other domains, there are some features of general-purpose AI that appear to create distinctive difficulties.

Several technical features of general-purpose AI make risk management in this domain particularly difficult. They include, among others:

- **The range of possible uses and use contexts for general-purpose AI systems is unusually broad.** For example, the same system may be used to provide medical advice, analyse computer code for vulnerabilities, and generate photos. This increases the difficulty of comprehensively anticipating relevant use cases, identifying risks, or testing how systems will behave in relevant real-world circumstances.
- **Developers still understand little about how their general-purpose AI models operate.** This lack of understanding makes it more difficult both to predict behavioural issues and to explain and resolve known issues once they are observed. Understanding remains elusive mainly because general-purpose AI models are not programmed in the traditional sense.

Instead, they are trained: AI developers set up a training process that involves a large volume of data, and the outcome of that training process is the general-purpose AI model. The inner workings of these models are largely inscrutable, including to the model developers. Model explanation and 'interpretability' techniques can improve researchers' and developers' understanding of how general-purpose AI models operate, but, despite recent progress, this research remains nascent.

- **Increasingly capable AI agents – general-purpose AI systems that can autonomously act, plan, and delegate to achieve goals – will likely present new, significant challenges for risk management.** AI agents typically work towards goals autonomously by using general software such as web browsers and programming tools. Currently, most are not yet reliable enough for widespread use, but companies are making large efforts to build more capable and reliable AI agents and have made progress in recent months. AI agents will likely become increasingly useful, but may also exacerbate a number of the risks discussed in this report and introduce additional difficulties for risk management. Examples of such potential new challenges include the possibility that users might not always know what their own AI agents are doing, the potential for AI agents to operate outside of anyone's control, the potential for attackers to 'hijack' agents, and the potential for AI-to-AI interactions to create complex new risks. Approaches for managing risks associated with agents are only beginning to be developed.

Besides technical factors, several economic, political, and other societal factors make risk management in the field of general-purpose AI particularly difficult.

- **The pace of advancement in general-purpose AI creates an 'evidence dilemma' for decision-makers.**[†] Rapid capability advancement makes it possible for some risks to emerge in leaps; for example, the risk of academic cheating using general-purpose AI shifted from negligible to widespread within a year. The more quickly a risk emerges, the more difficult it is to manage the risk reactively and the more valuable preparation becomes. However, so long as evidence for a risk remains incomplete, decision-makers also cannot know for sure whether the risk will emerge or perhaps even has already emerged. This creates a trade-off: implementing pre-emptive or early mitigation measures might prove unnecessary, but waiting for conclusive evidence could leave society vulnerable to risks that emerge rapidly. Companies and governments are developing early warning systems and risk management frameworks that may reduce this dilemma. Some of these trigger specific mitigation measures when there is new evidence of risks, while others require developers to provide evidence of safety before releasing a new model.
- **There is an information gap between what AI companies know about their AI systems and what governments and non-industry researchers know.** Companies often share only limited information about their general-purpose AI systems, especially in the period before they are widely released. Companies cite a mixture of commercial concerns and safety concerns as

[†] Please refer to the Chair's update on the latest AI advances after the writing of this report.

reasons to limit information sharing. However, this information gap also makes it more challenging for other actors to participate effectively in risk management, especially for emerging risks.

- **Both AI companies and governments often face strong competitive pressure, which may lead them to deprioritise risk management.** In some circumstances, competitive pressure may incentivise companies to invest less time or other resources into risk management than they otherwise would. Similarly, governments may invest less in policies to support risk management in cases where they perceive trade-offs between international competition and risk reduction.

Nonetheless, there are various techniques and frameworks for managing risks from general-purpose AI that companies can employ and regulators can require. These include methods for identifying and assessing risks, as well as methods for mitigating and monitoring them.

- **Assessing general-purpose AI systems for risks is an integral part of risk management, but existing risk assessments are severely limited**. Existing evaluations of general-purpose AI risk mainly rely on 'spot checks', i.e. testing the behaviour of a general-purpose AI in a set of specific situations. This can help surface potential hazards before deploying a model. However, existing tests often miss hazards and overestimate or underestimate general-purpose AI capabilities and risks, because test conditions differ from the real world.
- **For risk identification and assessment to be effective, evaluators need substantial expertise, resources, and sufficient access to relevant information.** Rigorous risk assessment in the context of general-purpose AI requires combining multiple evaluation approaches. These range from technical analyses of the models and systems themselves to evaluations of possible risks from certain use patterns. Evaluators need substantial expertise to conduct such evaluations correctly. For comprehensive risk assessments, they often also need more time, more direct access to the models and their training data, and more information about the technical methodologies used than the companies developing general-purpose AI typically provide.
- **There has been progress in training general-purpose AI models to function more safely, but no current method can reliably prevent even overtly unsafe outputs.** For example, a technique called 'adversarial training' involves deliberately exposing AI models to examples designed to make them fail or misbehave during training, aiming to build resistance to such cases. However, adversaries can still find new ways ('attacks') to circumvent these safeguards with low to moderate effort. In addition, recent evidence suggests that current training methods – which rely heavily on imperfect human feedback – may inadvertently incentivise models to mislead humans on difficult questions by making errors harder to spot. Improving the quantity and quality of this feedback is an avenue for progress, though nascent training techniques using AI to detect misleading behaviour also show promise.
- **Monitoring – identifying risks and evaluating performance once a model is already in use – and various interventions to prevent harmful actions can improve the safety of a general-purpose AI after it is deployed to users.** Current tools can detect AI-generated

content, track system performance, and identify potentially harmful inputs/outputs, though moderately skilled users can often circumvent these safeguards. Several layers of defence that combine technical monitoring and intervention capabilities with human oversight improve safety but can introduce costs and delays. In the future, hardware-enabled mechanisms could help customers and regulators to monitor general-purpose AI systems more effectively during deployment and potentially help verify agreements across borders, but reliable mechanisms of this kind do not yet exist.

- **Multiple methods exist across the AI lifecycle to safeguard privacy.** These include removing sensitive information from training data, model training approaches that control how much information is learned from data (such as 'differential privacy' approaches), and techniques for using AI with sensitive data that make it hard to recover the data (such as 'confidential computing' and other privacy-enhancing technologies). Many privacy-enhancing methods from other research fields are not yet applicable to general-purpose AI systems due to the computational requirements of AI systems. In recent months, privacy protection methods have expanded to address AI's growing use in sensitive domains including smartphone assistants, AI agents, always-listening voice assistants, and use in healthcare or legal practice.

Since the publication of the Interim Report, researchers have made some further progress towards being able to explain why a general-purpose AI model has produced a given output. Being able to explain AI decisions could help manage risks from malfunctions ranging from bias and factual inaccuracy to loss of control. In addition, there have been growing efforts to standardise assessment and mitigation approaches around the world.

Conclusion: A wide range of trajectories for the future of general-purpose AI are possible, and much will depend on how societies and governments act

The future of general-purpose AI is uncertain, with a wide range of trajectories appearing possible even in the near future, including both very positive and very negative outcomes. But nothing about the future of general-purpose AI is inevitable. How general-purpose AI gets developed and by whom, which problems it gets designed to solve, whether societies will be able to reap general-purpose AI's full economic potential, who benefits from it, the types of risks we expose ourselves to, and how much we invest into research to manage risks – these and many other questions depend on the choices that societies and governments make today and in the future to shape the development of general-purpose AI.

To help facilitate constructive discussion about these decisions, this report provides an overview of the current state of scientific research and discussion on managing the risks of general-purpose AI. The stakes are high. We look forward to continuing this effort.