

Interim international scientific report on advanced AI safety, chaired by Bengio, is out

Source copy kept by wheresthe.ai for the story "Interim international scientific report on advanced AI safety, chaired by Bengio, is out" (<https://wheresthe.ai/d/intl-ai-safety-report-interim-2024-05/>).

Source: <https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai>

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

SHA-256 of the saved text: e0b393b91500a8e256dfacc427b0843df55255e068fd2ec1a251b402e68ace53

Documents the source offers: https://assets.publishing.service.gov.uk/media/6716673b96def6d27a4c9b24/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf; https://assets.publishing.service.gov.uk/media/666ac0732f424458f904f308/international_scientific_report_on_the_safety_of_advanced_ai_executive_summary_arabic.pdf; https://assets.publishing.service.gov.uk/media/666ac68ca8b7ec4fae43d0f9/international_scientific_report_on_the_safety_of_advanced_ai_executive_summary_chinese.pdf

Research and analysis

International Scientific Report on the Safety of Advanced AI: interim report

An interim report on the science of advanced Artificial Intelligence (AI) safety.

Updates to this page

From:

Cabinet Office , AI Security Institute and Department for Science, Innovation and Technology

Published: 17 May 2024 Last updated:

22 October 2025

Show all updates

22 October 2025 Updated with a link to the external website of the International AI Safety Report: <https://internationalaisafetyreport.org/>.

17 May 2024 First published.

Get emails about this page

Documents

International scientific report on the safety of advanced AI: interim report

Ref: DSIT research paper series number 2024/009

PDF , 2.85 MB , 132 pages

This file may not be suitable for users of assistive technology.

Request an accessible format

If you use assistive technology (such as a screen reader) and need a version of this document in a more accessible format, please email alt.formats@dsit.gov.uk . Please tell us what format you need. It will help us if you say what assistive technology you use.

International scientific report on the safety of advanced AI: interim report

Ref: DSIT research paper series number 2024/009

HTML

International scientific report on the safety of advanced AI: executive summary (Arabic)

PDF , 207 KB , 8 pages

This file may not be suitable for users of assistive technology.

Request an accessible format

If you use assistive technology (such as a screen reader) and need a version of this document in a more accessible format, please email alt.formats@dsit.gov.uk . Please tell us what format you need. It will help us if you say what assistive technology you use.

International scientific report on the safety of advanced AI: executive summary (Chinese)

PDF , 470 KB , 7 pages

This file may not be suitable for users of assistive technology.

Request an accessible format

If you use assistive technology (such as a screen reader) and need a version of this document in a more accessible format, please email alt.formats@dsit.gov.uk . Please tell us what format you need. It will help us if you say what assistive technology you use.

International scientific report on the safety of advanced AI: executive summary (French)

PDF , 193 KB , 9 pages

This file may not be suitable for users of assistive technology.

Request an accessible format

If you use assistive technology (such as a screen reader) and need a version of this document in a more accessible format, please email alt.formats@dsit.gov.uk . Please tell us what format you need. It will help us if you say what assistive technology you use.

International scientific report on the safety of advanced AI: executive summary (Russian)

PDF , 239 KB , 10 pages

This file may not be suitable for users of assistive technology.

Request an accessible format

If you use assistive technology (such as a screen reader) and need a version of this document in a more accessible format, please email alt.formats@dsit.gov.uk . Please tell us what format you need. It will help us if you say what assistive technology you use.

International scientific report on the safety of advanced AI: executive summary (Spanish)

PDF , 190 KB , 9 pages

This file may not be suitable for users of assistive technology.

Request an accessible format

If you use assistive technology (such as a screen reader) and need a version of this document in a more accessible format, please email alt.formats@dsit.gov.uk . Please tell us what format you need. It will help us if you say what assistive technology you use.

Details

A full report - International AI Safety Report 2025 , was published in January 2025 ahead of the AI Action Summit in France.

Further information and updates are available at <https://internationalaisafetyreport.org/> .

The International Scientific Report on the Safety of Advanced AI interim report sets out a science-based understanding of the safety of advanced AI systems. The independent, international, and inclusive report is a landmark moment of international collaboration. It marks the first time the international community has come together to support efforts to build a shared scientific and evidence-based understanding of frontier AI risks.

The intention to create such a report was announced at the AI Safety Summit in November 2023 .

The interim report restricts its focus to a summary of the evidence on general-purpose AI , which have advanced rapidly in recent years. The report synthesises the evidence base on the capabilities of, and risks from, general-purpose AI and evaluates technical methods for assessing and mitigating them.

The interim report highlights several key takeaways, including:

General-purpose AI can be used to advance the public interest, leading to enhanced wellbeing, prosperity, and scientific discoveries.

According to many metrics, the capabilities of general-purpose AI are advancing rapidly. Whether there has been significant progress on fundamental challenges such as causal reasoning is debated among researchers.

Experts disagree on the expected pace of future progress of general-purpose AI capabilities, variously supporting the possibility of slow, rapid, or extremely rapid progress.

There is limited understanding of the capabilities and inner workings of general-purpose AI systems. Improving our understanding should be a priority.

Like all powerful technologies, current and future general-purpose AI can be used to cause harm. For example, malicious actors can use AI for large-scale disinformation and influence operations, fraud, and scams.

Malfunctioning general-purpose AI can also cause harm, for instance through biased decisions with respect to protected characteristics like race, gender, culture, age, and disability.

Future advances in general-purpose AI could pose systemic risks, including labour market disruption, and economic power inequalities. Experts have different views on the risk of humanity losing control over AI in a way that could result in catastrophic outcomes.

Several technical methods (including benchmarking, red-teaming and auditing training data) can help to mitigate risks, though all current methods have limitations, and improvements are required.

The future of AI is uncertain, with a wide range of scenarios appearing possible. The decisions of societies and governments will significantly impact its future.

The report underlines the need for continuing collaborative international efforts to research and share knowledge about these rapidly evolving technologies. The approach taken was deliberately inclusive of different views and perspectives, and areas of uncertainty, consensus or dissent are highlighted, promoting transparency.

Background information

The report was commissioned by the UK government and chaired by Yoshua Bengio, a Turing Award-winning AI academic and member of the UN's Scientific Advisory Board. The work was overseen by an international Expert Advisory Panel made up of 30 countries including the UK and nominees from nations who were invited to the AI Safety Summit at Bletchley Park in 2023, as well as representatives of the European Union and the United Nations.

The report aim's to drive a shared, science-based, up-to-date understanding of the safety of advanced AI systems, and to develop that understanding over time. To do so, the report brings together world-leading AI countries and the best global AI expertise to analyse existing scientific research on AI capabilities and risks.

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

[Sign up for emails or print this page](#)

[Get emails about this page](#)

[Print this page](#)