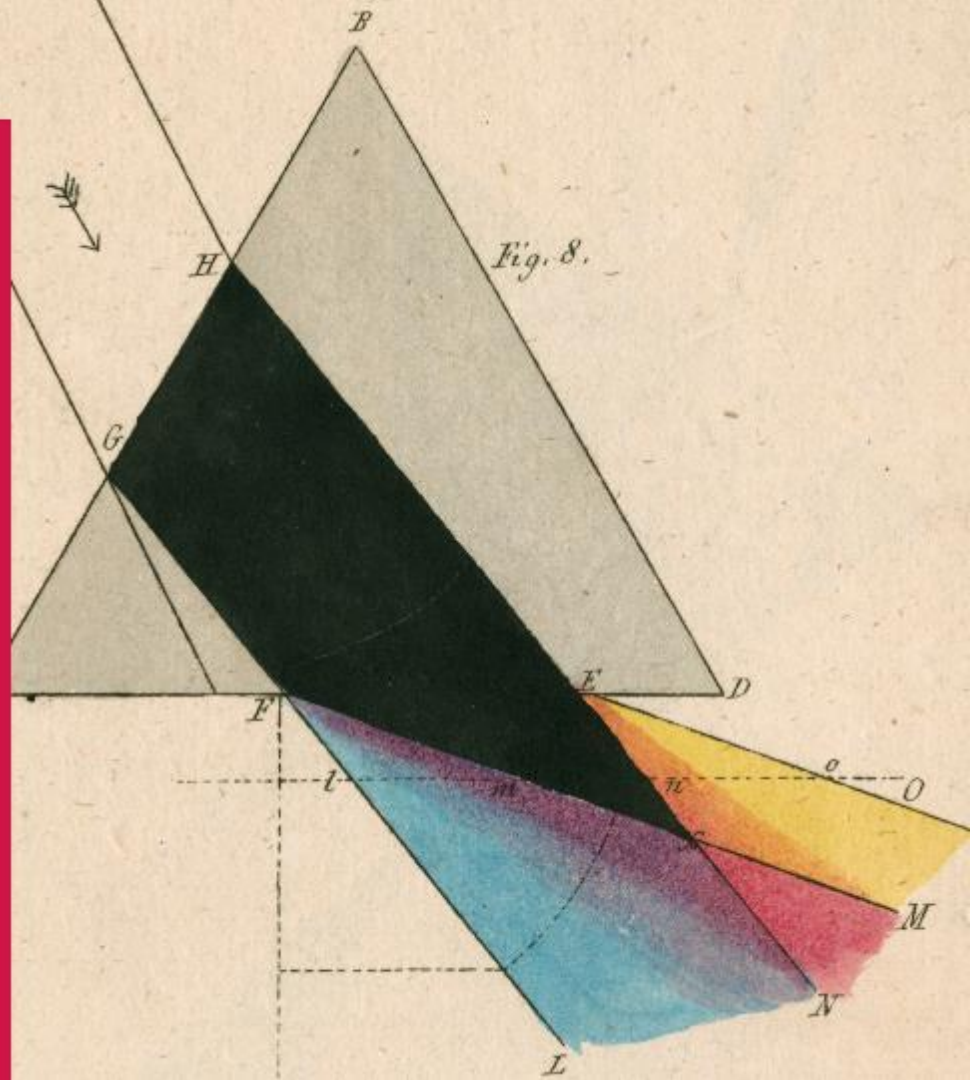


AI in research: Lessons from *Science in the age of AI*

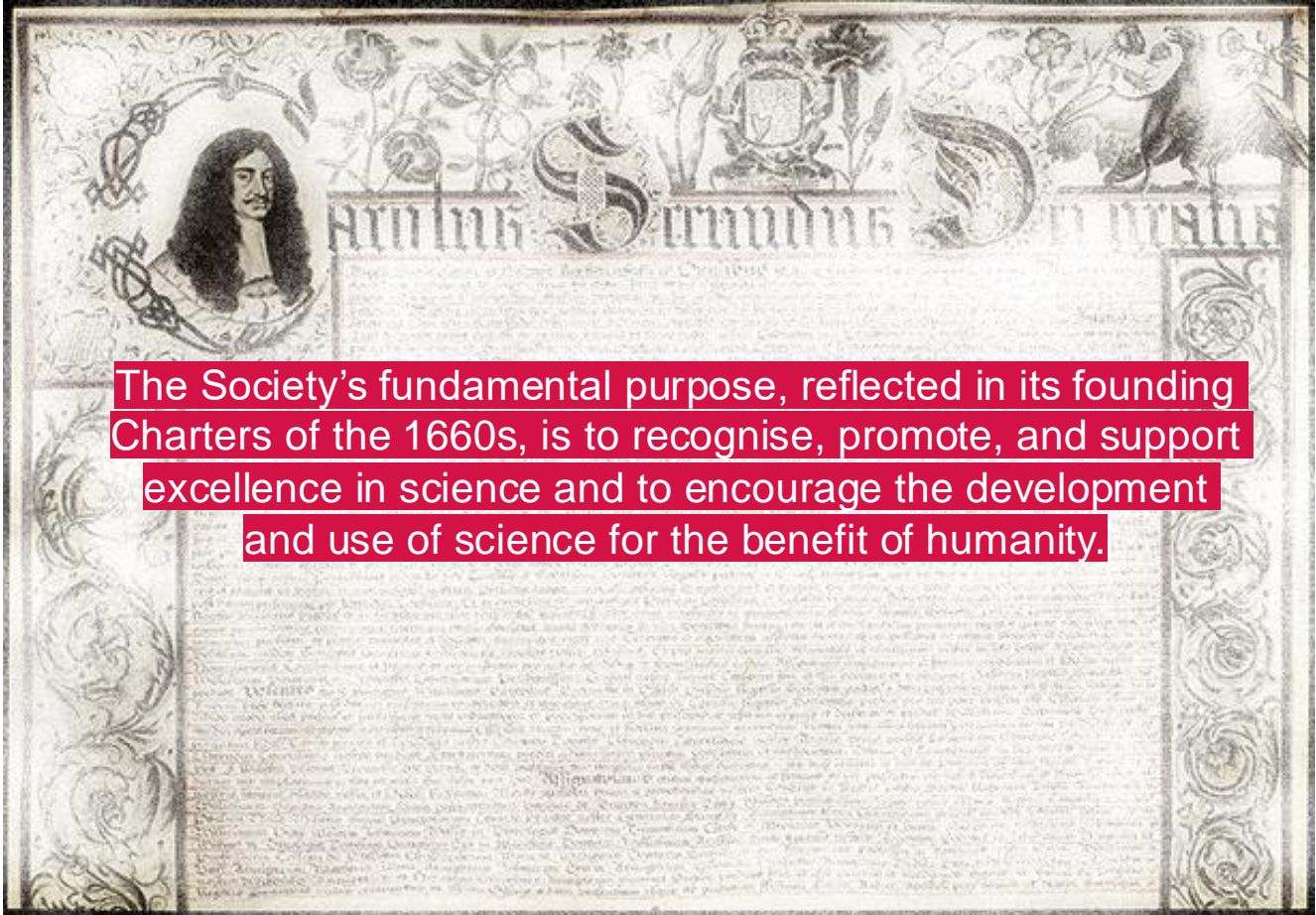
Areeq Chowdhury
Head of Policy, Data and Digital Technologies
areeq.chowdhury@royalsociety.org

UK Research Integrity Office, 9 October 2024





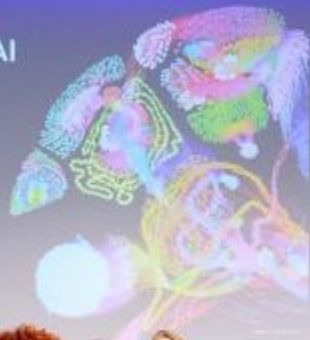
Our mission.



The Society's fundamental purpose, reflected in its founding Charters of the 1660s, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity.

Science in the age of AI

How artificial intelligence
is changing the nature and
method of scientific research



THE
ROYAL
SOCETY



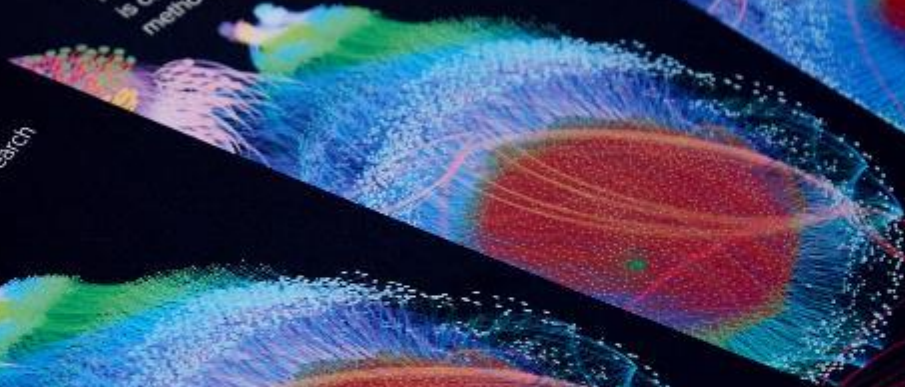
Science in the age of AI

How artificial intelligence
is changing the nature and
method of scientific research



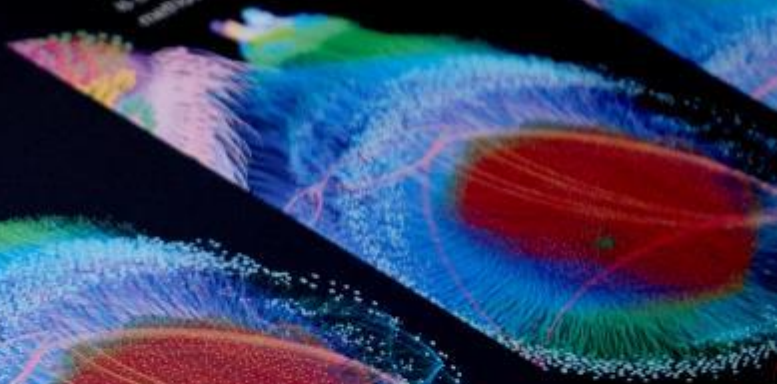
Science in the age of AI

How artificial intelligence
is changing the nature and
method of scientific research



the

How artificial intelligence
is changing the nature
method of scientific research



SCIENCE IN THE AGE OF AI

How artificial intelligence is changing the nature and method of scientific research

Chair



Professor Alison Noble CBE FRS FEng – Vice President of the Royal Society and Technikos Professor of Biomedical Engineering, University of Oxford.

Members

Professor Paul Beasley – Head of Research and Development, Siemens.

Dr Peter Dayan FRS – Director, Max Planck Institute for Biological Cybernetics.

Professor Sabina Leonelli – Professor of Philosophy and History of Science, University of Exeter.

Alistair Nolan – Senior Policy Analyst, Organisation for Economic Co-operation and Development.

Dr Philip Quinlan – Director of Health Informatics, University of Nottingham.

Professor Abigail Sellen FRS – Distinguished Scientist and Lab Director, Microsoft Research.

Professor Rossi Setchi – Professor in High Value Manufacturing, Cardiff University.

Kelly Vere – Director of Technical Strategy, University of Nottingham



Chapters.

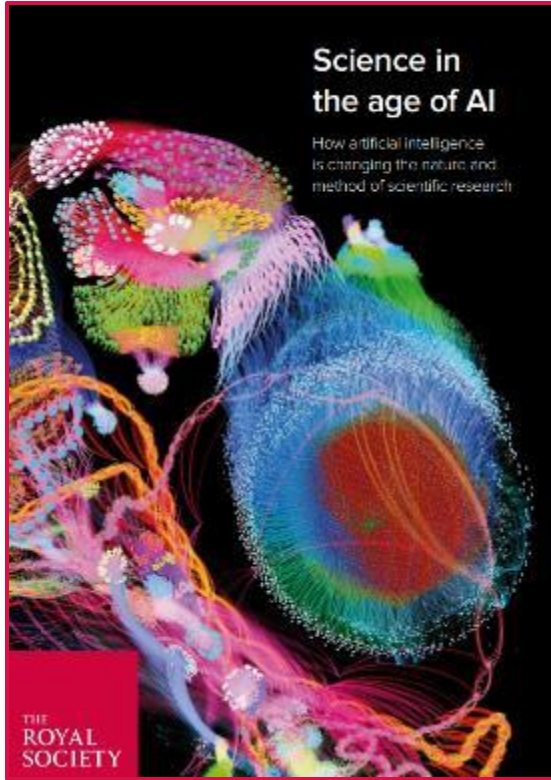
- ▶ Overview: How AI is transforming scientific research
- ▶ Research integrity and trustworthiness
- ▶ Research skills and interdisciplinarity
- ▶ Research, innovation and the role of the private sector
- ▶ Research ethics and AI safety
- ▶ 3 case studies
 - ▶ Material science
 - ▶ Climate science
 - ▶ Rare disease research

Methodology.

- ▶ 30+ interviews
- ▶ 5 roundtables
 - ▶ Immersive technologies (2022)
 - ▶ Reproducibility (May 2023)
 - ▶ AI and climate science (June 2023)
 - ▶ Interdisciplinarity (July 2023)
 - ▶ LLMs and science (July 2023)
- ▶ 2 AI safety workshops (October 2023)
 - ▶ Horizon scanning AI safety risks in science
 - ▶ Red teaming on AI-generated disinformation
- ▶ 2 international workshops (Sept – Nov 23)
 - ▶ UK-US Researcher Access to Data Forum, Washington DC
 - ▶ RS-CAS workshop on AI Ethics, Beijing
- ▶ 3 commissioned studies
 - ▶ Taxonomy of AI technologies
 - ▶ Patent landscape of AI technologies
 - ▶ Historical review



Areas for action.



1

Enhance access to essential AI infrastructures and tools

2

Trust in the quality of AI-based scientific outputs

3

Ensure safe and ethical use of AI in scientific research

Recommendations.

1. Governments, research funders and AI developers should improve access to essential AI infrastructures.
2. Funders and AI developers should prioritise accessibility and usability of AI tools developed for scientific research.
3. Research funders and scientific communities should ensure AI-based research meets open science principles and practices to facilitate AI's benefits in science.
4. Scientific communities should build the capacity to oversee AI systems used in science and ensure their ethical use for the public good.



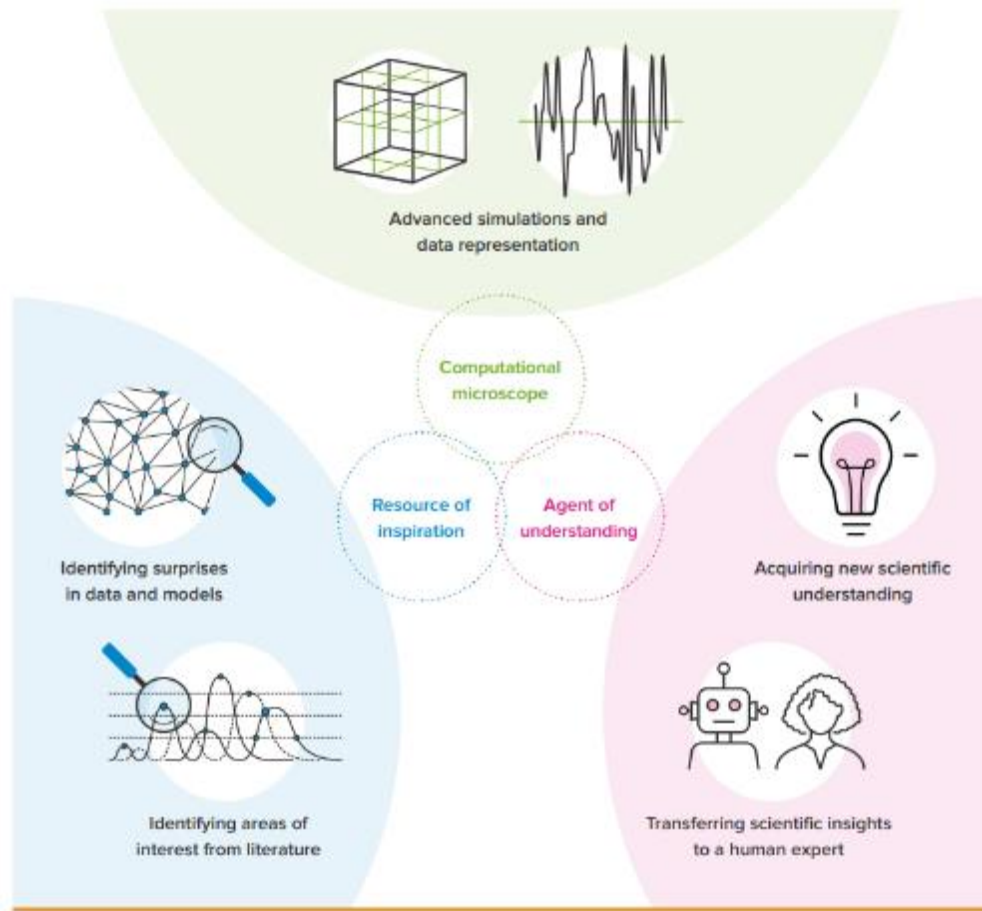


Key findings.

1. AI applications can be found across all STEM fields.
2. High quality data is foundational for AI applications.
3. China, USA, Japan, and South Korea are dominating in terms of patent applications related to AI for science.
4. Companies such as Alphabet, Siemens, IBM, and Samsung appear to exhibit considerable influence.
5. The black-box, and potentially proprietary, nature of AI tools is limiting the reproducibility of AI-based research.
6. Interdisciplinary collaboration is essential to bridge skills gaps.
7. Generative AI tools hold promise for expediting routine scientific tasks.

FIGURE 1

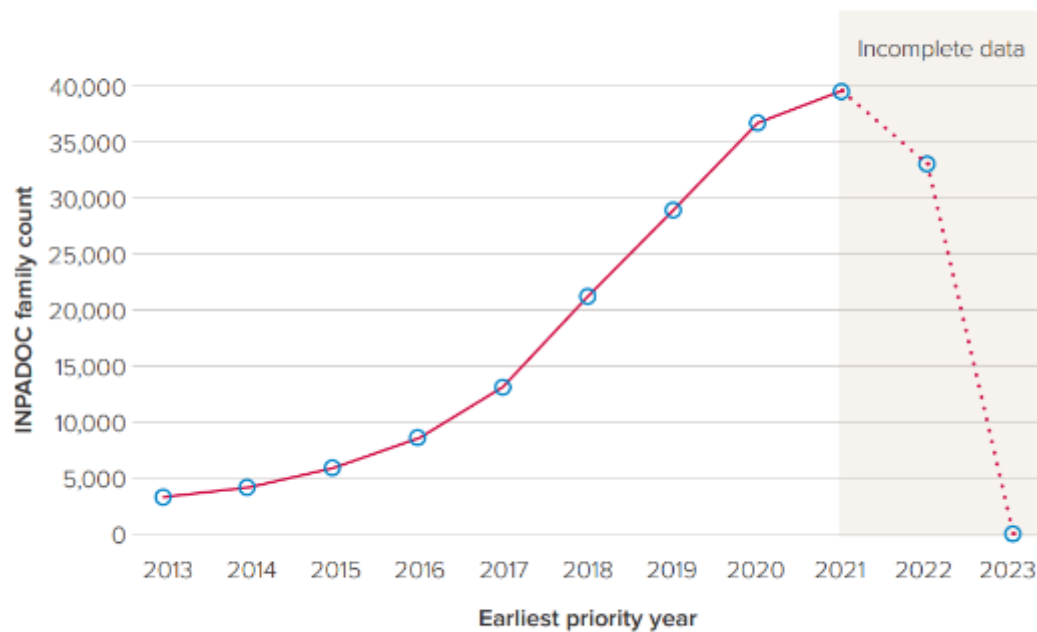
Reproduction of a visualisation of the three general roles of AI for scientific research as either a computational microscope, resource of human inspiration, or an agent of understanding²⁴.



See: Krenn, M et al 2022. *On Scientific Understanding with Artificial Intelligence*.

FIGURE 2

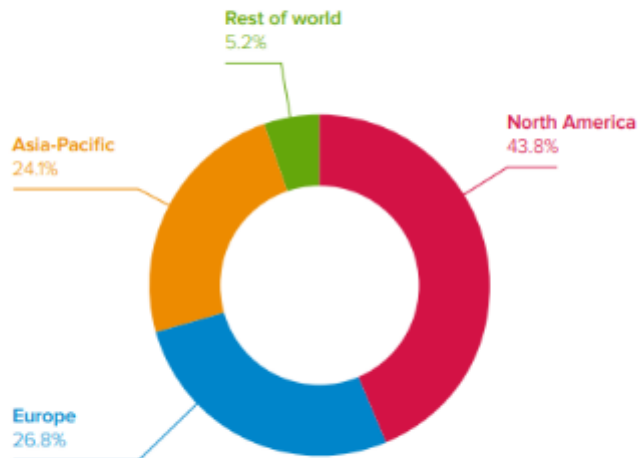
Patent filing trends of AI-related technological inventions in the last 10 years



(Data for 2021 – 2023 is not complete given the 18-month delay from the priority filing date and the date of publication).

FIGURE 4

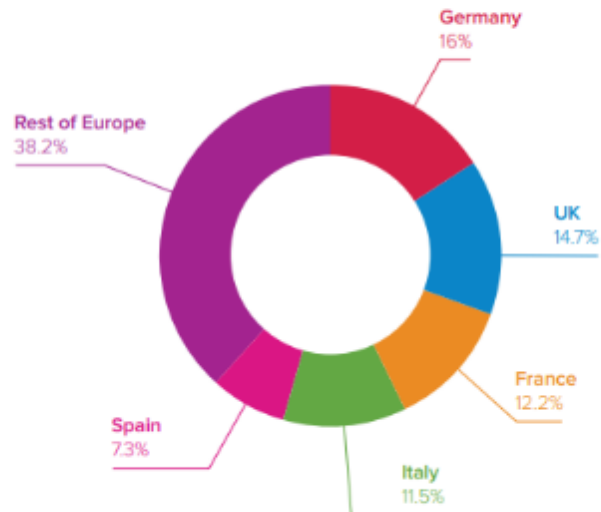
Global Market Shares of Machine Learning in the Life Sciences, by Region, 2021 (%)



Source: BCC Research.

FIGURE 5

European Market Shares of Machine Learning in the Life Sciences, by Country, 2021 (%)



Source: BCC Research.

“

It is hardly possible to imagine higher stakes than these for the world of science. The future existence and social role [of science] seem to hinge on the ability of researchers and scientific institutions to respond to the [reproducibility] crisis, thus averting a complete loss of trust in scientific expertise by civil society.

”

Explainability and interpretability

Explainability and interpretability refer to information that allows users to understand how an AI system works and the reasoning behind its outputs¹⁷⁸. For example, in ML interpretability methods can offer information into 'how a model works' while explainability answers why certain conclusions are reached or "what else can this model tell me?"¹⁷⁹.

As set out in the Royal Society's 2019 report, *Explainable AI: The basics*¹⁸⁰, ensuring explainability and interpretability in science can have the following benefits for trustworthiness:

- Helps researchers better understand the insights and patterns that come from the use of complex machine learning models and large datasets.
- Enhances the potential for scientists to draw insights from AI systems to reveal potential new scientific breakthroughs or discoveries¹⁸¹.
- Improves reproducibility by enabling third parties to scrutinise the model, as well as identify and correct errors.
- Improves transferability and assessment of whether models could be suitable across disciplines or contexts.
- Improves accountability and ensures scientists can offer justification behind the use of ML models¹⁸².
- In the case of science-based applications that affect the public – from health to public policy – explainability can ensure policy makers and regulators can provide oversight and prevent harms caused by erroneous predictions or models¹⁸³.

TABLE 1

Barriers to reproducibility and examples.

Barrier to reproducibility	Examples
Misconceptions and assumptions about ML²⁰⁰	<ul style="list-style-type: none"> • An underlying assumption that machine learning (ML) models are inherently reproducible due to their reliance on computation. • Overreliance on ML-based outputs and questionable uses of statistical techniques to smoothen bias or exclude uncomfortable or inconvenient results.
Computational or environmental conditions	<ul style="list-style-type: none"> • Different hardware and software environments may yield different results. • Reproducibility at scale implies having access to computation capacity that enables researchers to validate complex machine learning models²⁰¹. • Private sector companies are better resourced than academia and can afford to train and validate larger models (eg OpenAI's GPT-4) while researchers in other sectors cannot²⁰².
Documentation and transparency practices	<ul style="list-style-type: none"> • Insufficient or incomplete documentation around research methods, code, data, or computational environments. • The growing development and adoption of less transparent, proprietary models. • Lack of discipline-specific documentation that addresses barriers faced across fields, applications, and research contexts (eg healthcare-specific documentation that tackles reproducibility guidelines for disease treatment and diagnosis research). • Insufficient efforts to make documentation accessible to scientists from different backgrounds and with diverse levels of technical expertise.
Skills, training and capacity	<ul style="list-style-type: none"> • Lack of clarity regarding who is responsible for different stages of the workflow and few resources to incorporate reproducibility work. • Lack of training for new ML users and insufficient guidelines on the limitations of different models and the appropriateness of different techniques for field-specific applications. • Lack of tools for non-ML experts to follow reproducibility guidelines and identify limitations of models. • Lack of mechanisms that facilitate interdisciplinary collaboration between scientists who do not have a technical background in AI and computer or data scientists who carry expertise to input data, identify errors, and validate experiments.
Incentives and research culture	<ul style="list-style-type: none"> • Few career progression opportunities in academia for roles needed to advance open and reproducible research (eg data curation and wrangling; research data management; data stewardship; research managers). • No incentives to publish errors in ML-based research (failed results) or remedies. • Narrow view of what outputs are worthy of publishing (eg data, models) and limited rewards for conducting open science practices and publishing reproducibility reports. • No specific incentives to encourage the use and development of human-interpretable models when possible²⁰³.

AI safety.





TECH ARTIFICIAL INTELLIGENCE

The Scientists Breaking AI to Make It Safer



Experts in climate science and disease tried to coax misinformation out of AI programs at an event at London's Royal Society. Courtesy Royal Society

BY BILLY FERRIGO

OCTOBER 26, 2023 2:02 PM EDT

In an ornate room lined with marble busts of famous scientists, around 40 experts in climate science and disease were hunched over their laptops yesterday (Oct. 25), coaxing a powerful AI system into generating misinformation.

By the end of the day, attendees had managed to overcome the guardrails on the AI system—Meta's Llama 2—and got it to argue that ducks could absorb air pollution, to say that garlic and “miraculous herbs” could help prevent COVID-

Red teaming large language models (LLMs) for resilience to scientific disinformation

Summary note of an event held on 25 October 2023

Background

The Royal Society and Humane Intelligence co-hosted a red teaming event in the run-up to the 2023 Global AI Safety Summit (Bletchley, UK). The red teaming event brought together 40 health and climate postgraduate students with the objective to scrutinise and bring attention to potential vulnerabilities in large language models (LLMs). Since the viral release of ChatGPT in late 2022, LLMs have seen increased uptake by both scientists and the general public, prompting concerns of a rising AI-driven infodemic¹. This event invited participants to explore the nature of these potential harms and contribute to discussions related to the use of generative AI in the production of scientific misinformation and disinformation².

The event took place on 25 October 2023 and was part of the Science & AI Safety series of events hosted at the Royal Society which explored the risks associated with the use of AI in scientific activities. Building on the report *The online infodemic environment: Understanding how the internet shapes people's engagement with scientific information*, published in January 2022, the activity aimed to explore AI-generated scientific disinformation, and provide insights on the efficacy of guardrails to prevent its production and dissemination. An additional objective was to understand the opportunities and limitations of involving scientists in red teaming efforts.

What is red teaming?

Red teaming is a socio-technical evaluation method, in which a group of people is authorised to act as an adversary (the ‘red team’), emulating attacks and exploiting the vulnerabilities of a system. Red teaming techniques are related to so-called ‘jailbreaking’ methods, and involve crafting prompts to bypass safety features and eliciting text or code generation that could be harmful or otherwise undesirable (eg mislead readers, hate speech, or their automated dissemination).

While red teaming is rooted in cyber security practices, it has wider implications, for example, in ‘stress testing’ new technologies like AI applications. It has also emerged in AI policy and governance discussions as a promising approach for identifying the potential harms of LLMs³. In this context, red teaming could be used to inform development, testing, and validation strategies for AI risk mitigation.

1. Large language models, or LLMs, refers to a type of artificial intelligence system designed to understand and generate text after being trained on vast amounts of training data. They can produce and produce text across various topics and styles, they can also perform a wide range of tasks, language processing tasks, such as text generation, language translation and sentiment analysis.
2. De Angelis, L. et al. 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health* 11:1047456 (2023).
3. The Royal Society. 2023. Generative AI: current provisions and a public service roadmap. <https://royalsocietypublishing.org/journal/rsos/10/2302023/generative-ai-current-provisions-and-a-public-service-roadmap>
4. Royal Society. 2022. *The online infodemic environment: Understanding how the internet shapes people's engagement with scientific information*. <https://royalsocietypublishing.org/journal/rsos/10/2202022/the-online-infodemic-environment> (accessed 15 February 2024).
5. Several governments' interest in red teaming LLMs for AI risk mitigation reflects the UK and United States, see UK Government. Introducing the AI Safety Institute. <https://www.gov.uk/government/news/safety-institute-opens> (accessed 15 February 2024); see also The White House. 2023. Red teaming large language models to identify and fix AI risks. <https://www.whitehouse.gov/briefings-statements/2023/06/29/red-teaming-large-language-models-to-identify-and-fix-ai-risks/> (accessed 15 February 2024).

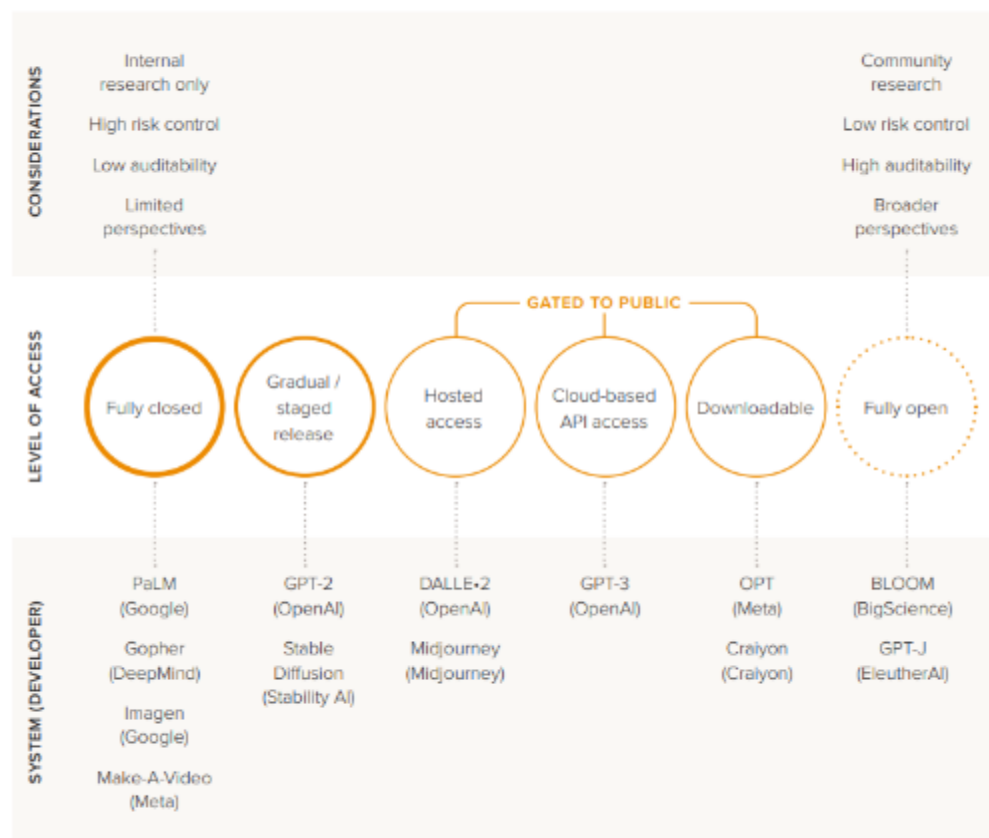


[News >](#)

Making AI more open could accelerate research and tech transfer

Combining artificial Intelligence (AI) and open science could accelerate scientific discovery, redefine the boundaries of scientific research and democratise access to knowledge, suggested participants in a symposium on 6 June co-hosted by UNESCO and the Royal Society in the UK, which also featured the launch of the latter's report on Science in the Age of AI.

Reproduction of the Gradient of System Access developed by Hugging Face





“

Everybody wants the sparkly fountain, but very few people are thinking of the boring plumbing system underneath it.

”



For more information

royalsociety.org/science-in-the-age-of-ai
areeq.chowdhury@royalsociety.org