Science in the age of AI

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

THE ROYAL SOCIETY

Science in the age of AI: How artificial intelligence is changing the nature and method of scientific research
Issued: October 2024 DES8836_5
ISBN: 978-1-78252-712-1

This report and other project outputs can be viewed at:
royalsociety.org/science-in-the-age-of-ai

Cover image: Computational graph developed by Graphcore, visualising what the algorithms in a machine learning model look like when they are in action. This particular graph is a mapping of the deep learning tool ResNet18. © Graphcore.

Contents

Foreword

Image: Professor Alison Noble FRS.

With the growing availability of large datasets, new algorithmic techniques and increased computing power, artificial intelligence (AI) is becoming an established tool used by researchers across scientific fields.

Now more than ever, we need to understand the extent of the transformative impact of AI on science and what scientific communities need to do to fully harness its benefits.

This report, *Science in the age of AI*, explores this topic. Building on the experiences of more than 100 scientists who have incorporated AI into their workflows, it delves into how AI technologies, such as deep learning or large language models, are transforming the nature and methods of scientific inquiry. It also explores how notions of research integrity, research skills and research ethics are inevitably changing – and what the implications are for the future of science and scientists.

New opportunities are emerging. The case studies in this report demonstrate that AI is enhancing the efficiency, accuracy, and creativity of scientists. Across multiple fields, the application of AI is breaking new ground by facilitating, for example, the discovery of rare diseases or enabling the development of more sustainable materials.

Playing the role of tutor, peer or assistant, scientists are using AI applications to perform tasks at a pace and scale previously unattainable. There is much excitement around the synergy between human intelligence and AI and how this partnership is leading to scientific advancements. However, to ensure robustness and mitigate harms, human judgement and expertise will continue to be of utmost importance.

The rapid uptake of AI in science has also presented challenges related to its safe and rigorous use. A growing body of irreproducible studies are raising concerns regarding the robustness of AI-based discoveries. The black-box and non-transparent nature of AI systems creates challenges for verification and external scrutiny. Furthermore, its widespread but inequitable adoption raises ethical questions regarding its environmental and societal impact. Yet, ongoing advancements in making AI systems more transparent and ethically aligned hold the promise of overcoming these challenges.

In this regard, the report calls for a balanced approach that celebrates the potential of AI in science while not losing sight of the challenges that still need to be overcome. The recommendations offer a pathway that leverages open science principles to enable reliable AI-driven scientific contributions, while creating opportunities for resource sharing and collaboration. They also call for policies and practices that recognise the links between science and society, emphasising the need for ethical AI, equitable access to its benefits, and the importance of keeping public trust in scientific research.

While it's clear that AI can significantly aid scientific advancement, the goal remains to ensure these breakthroughs benefit humanity and the planet. We hope this report inspires actors across the scientific ecosystem to engage with the recommendations and work towards a future where we can realise the potential of AI to transform science and benefit our collective wellbeing.

Professor Alison Noble CBE FREng FRS, Foreign Secretary of the Royal Society and Chair of the Royal Society Science in the Age of AI Working Group.

Executive summary

The unprecedented speed and scale of progress with artificial intelligence (AI) in recent years suggests society may be living through an inflection point. The virality of platforms such as ChatGPT and Midjourney, which can generate human-like text and image content, has accelerated public interest in the field and raised flags for policymakers who have concerns about how AI-based technologies may be integrated into wider society. Beyond this, comments made by prominent computer scientists and public figures regarding the risks AI poses to humanity have transformed the subject into a mainstream political issue. For scientific researchers, AI is not a novel topic and has been adopted in some form for decades. However, the increased investment, interest, and adoption within academic and industry-led research has led to a 'deep learning revolution'[1] that is transforming the landscape of scientific discovery.

Enabled by the advent of big data (for instance, large and heterogenous forms of data gathered from telescopes, satellites, and other advanced sensors), AI-based techniques are helping to identify new patterns and relationships in large datasets which would otherwise be too difficult to recognise. This offers substantial potential for scientific research and is encouraging scientists to adopt more complex techniques that outperform existing methods in their fields. The capability of AI tools to identify patterns from existing content and generate new predictions also allows scientists to run more accurate simulations and create synthetic data. These simulations, which draw data from lots of different sources (potentially in real time), can help decision-makers assess more accurately the efficacy of potential interventions and address pressing societal or environmental challenges.

The opportunities of AI for scientific research are highlighted throughout this report and explored in depth through three case studies on its application for climate science, material science, and rare disease diagnosis.

Alongside these opportunities, there are various challenges arising from the increased adoption of AI. These include reproducibility (in which other researchers cannot replicate experiments conducted using AI tools); interdisciplinarity (where limited collaboration between AI and non-AI disciplines can lead to a less rigorous uptake of AI across domains); and environmental costs (due to high energy consumption being required to operate large compute infrastructure). There are also growing barriers to the effective adoption of open science principles due to the black-box nature of AI systems and the limited transparency of commercial models that power AI-based research. Furthermore, the changing incentives across the scientific ecosystem may be increasing pressure on researchers to incorporate advanced AI techniques at the neglect of more conventional methodologies, or to be 'good at AI' rather than 'good at science'[2].

These challenges, and potential solutions, are detailed throughout this report in the chapters on research integrity; skills and interdisciplinarity; innovation and the private sector; and research ethics.

1 Sejnowski T. 2018 *The Deep Learning Revolution.* MIT press

2 Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See https://royalsociety.org/current-topics/ai-data/. (accessed 7 May 2024)

As an organisation that exists to promote the use of science for the benefit of humanity, this subject is of great importance to the Royal Society. This report, *Science in the Age of AI*, provides an overview of key issues to address for AI to positively transform the scientific endeavour. Its recommendations, when taken together, should ensure that the application of AI in scientific research is able to reach its full potential and help maintain public trust in science and the integrity of the scientific method.

This report has been guided by a working group of leading experts in AI and applied science and informed by a series of activities undertaken by the Royal Society. These include interviews with Fellows of the Royal Society; a global patent landscape analysis; a historical literature review; a commissioned taxonomy of AI for scientific applications; and several workshops on topics ranging from large language models to immersive technologies. These activities are listed in full in the appendix. In total, more than 100 leading scientific researchers from diverse disciplines contributed to this report.

While the report covers some of the critical areas related to the role of AI in scientific research, it is not comprehensive and does not cover, for example, the provision of high-performance computing infrastructure, the potential of artificial general intelligence, nor a detailed breakdown of the new skills required across industries and academia.

Further research questions are outlined below. The Society's two programmes of work on *Mathematical Futures*[3] and *Science 2040*[4] will explore, in more depth, relevant challenges related to skills and universities.

Key findings

- Beyond landmark cases like AlphaFold, AI applications can be found across all STEM fields, with a concentration in fields such as medicine, materials science, robotics, agriculture, genetics, and computer science. The most prominent AI techniques across STEM fields include artificial neural networks, deep learning, natural language processing and image recognition[5].
- High quality data is foundational for AI applications, but researchers face barriers related to the volume, heterogeneity, sensitivity, and bias of available data. The large volume of some scientific data (eg collected from telescopes and satellites) can total petabytes, making objectives such as data sharing and interoperability difficult to achieve. The heterogeneity of data collected from sensor data also presents difficulties for human annotation and standardisation, while the training of AI models on biased inputs can likely lead to biased outputs. Given these challenges, data curators and information managers are essential to maintain quality and address risks linked to artificial data generation, such as data fabrication, poisoning, or contamination.

3 The Royal Society. *Mathematical Futures* Programme. See https://royalsociety.org/-/media/policy/projects/maths-futures/mathematical-and-data-education-policy-report.pdf (accessed 23 April 2024)

4 The Royal Society. *Science 2040*. See https://royalsociety.org/news-resources/projects/science2040/ (accessed 23 April 2024)

5 Berman B, Chubb J, and Williams K, 2024. The use of artificial intelligence in science, technology, engineering, and medicine. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

- Industry and academic institutions are advancing AI innovation for scientific research[6]. The past decade has seen a surge in patent applications related to AI for science, with China, the United States, Japan, and South Korea dominating the number of patents filed in these territories. A review commissioned for this report suggests the valuation of the global AI market (as of 2022) is approximately £106.99 billion[7].
- China contributes approximately 62% of the patent landscape. Within Europe, the UK has the second largest share of AI patents related to life sciences after Germany, with academic institutions such as the University of Oxford, Imperial College, and Cambridge University featuring prominently among the top patent filers in the UK. Companies such as Alphabet, Siemens, IBM, and Samsung appear to exhibit considerable influence across scientific and engineering fields.
- The black-box, and potentially proprietary, nature of AI tools is limiting the reproducibility of AI-based research. Barriers such as insufficient documentation, limited access to essential infrastructures (eg code, data, and computing power) and a lack of understanding of how AI tools reach their conclusions (explainability) make it difficult for independent researchers to scrutinise, verify and replicate experiments. The significant potential to advance discoveries using complex deep learning models may also encourage scientists or funders to prioritise AI use over rigour. The adoption of open science principles and practices could help address these challenges and enhance scientific integrity[8].
- Interdisciplinary collaboration is essential to bridge skill gaps and optimise the benefits of AI in scientific research. By sharing knowledge and skills from each other's fields, collaboration between AI and domain subject experts (including researchers from the arts, humanities, and social sciences) can help produce more effective and accurate AI models. This is being prevented, however, by siloed research environments and an incentive structure that does not reward interdisciplinary collaboration in terms of contribution towards career progression.
- Generative AI tools can assist the advancement of scientific research. They hold promise for expediting routine scientific tasks, such as processing unstructured data, solving complex coding challenges, or supporting the multilingual translation of academic articles. In addition, there may be a place for text-generation models to be used for academic and non-academic written tasks, with potential implications for scholarly communications and research assessment. In response, funders and academic institutions are setting norms to prevent non-desirable uses[9,10].

6 Ahmed, N, Wahed, M, & Thompson, N. C. 2023. The growing influence of industry in AI research. *Science*, *379*(6635), 884-886. (DOI: 10.1126/science.ade2420)

7 IP Pragmatics, 2024 Artificial intelligence related inventions. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

8 UNESCO Recommendation on Open Science. 2021. See: https://www.unesco.org/en/legal-affairs/recommendation-open-science (accessed 6 February 2024)

9 Kaiser J. 2023. Funding agencies say no to AI peer review. *Science*. 14 July 2023 .See: https://www.science.org/content/article/science-funding-agencies-say-no-using-ai-peer-review (accessed 23 January 2024)

10 Harker J. 2023. Science Journals set new authorship guidelines for AI-generated text. National Institute of Environmental Health Sciences. See https://factor.niehs.nih.gov/2023/3/feature/2-artificial-intelligence-ethics. (accessed 23 January 2024)

Future research questions

The following topics emerged in research activities as key considerations for the future of AI in science:

1. **AI and computing infrastructures for science:** How can AI workloads be optimised to harness the full potential of heterogeneous computing infrastructures in scientific research, considering the diverse requirements of different scientific domains?
2. **AI and small data:** What are the implications of the growing use of AI for researchers in which only small data is available? How can AI techniques be effectively used to augment small datasets for training purposes? What trade-offs exist between model size reduction and preservation of performance when applied to small data scenarios?
3. **AI and inequities in the scientific system:** What barriers exist in providing equitable access to AI technologies in underrepresented communities? How can AI be used to broaden participation among scientific and expert communities, including underrepresented scholars and non-scientist publics?
4. **AI and intellectual property:** What inputs of AI systems (datasets, algorithms, or outputs) are crucial for intellectual property protection, and in what ways does it interact with the application of open science principles in science?
5. **AI and the future of skills for science:** How are the skill requirements in scientific research changing with the increasing integration of AI? What competencies will be essential for researchers in the future and what efforts are needed to promote AI literacy across diverse scientific disciplines?
6. **AI and the future of scholarly communication:** How is the landscape of scholarly and science communication evolving with the integration of AI technologies? How can AI be leveraged to improve knowledge translation, multilingualism, and multimodality in scholarly outputs?
7. **AI and environmental sustainability:** What role can AI play in promoting sustainable practices within the scientific community? How can AI algorithms be optimised to enhance the energy efficiency of environmental modelling, and contribute to sustainable practices in fields such as climate science, ecology, and environmental monitoring?
8. **AI standards and scientific research:** How can AI standards help address the challenges of reproducibility or interoperability in AI-based scientific research? How can the scientific community contribute to the establishment of AI standards?

Recommendations

AREA FOR ACTION: ENHANCE ACCESS TO ESSENTIAL AI INFRASTRUCTURES AND TOOLS

RECOMMENDATION 1

Governments, research funders and AI developers should improve access to essential AI infrastructures

Access to computing resources has been critical for major scientific breakthroughs, such as protein folding with AlphaFold. Despite this, compute power and data infrastructures for AI research are not equally accessible or distributed across research communities[11]. Scientists from diverse disciplines require access to infrastructure to adopt more complex AI techniques, process higher volume and types of data, and ensure quality in AI-based research.

Proposals to improve access have included institutions sponsoring access to supercomputing[12] and the establishment of regional hubs – akin to a CERN for AI[13]. Wider access can extend the benefits of AI to a greater number of disciplines, improve the competitiveness of non-industry researchers, and contribute towards more rigorous science by enabling reproducibility at scale. Expanding access to computing must also be informed by environmentally sustainable computational science (ESCS) best practices, including the measurement and reporting of environmental impacts[14].

Actions to enhance access to AI infrastructures and tools may include:

1. Funders, industry partners, and research institutions with computing facilities actively sharing essential AI infrastructures such as high-performance computing power and data resources.
2. Relevant stakeholders (eg government agencies, research institutions, industry, and international organisations) ensuring access to high-quality datasets and interoperable data infrastructures across sectors and regions. This could involve advancing access to sensitive data through privacy enhancing technologies and trusted research environments[15].
3. Research funders supporting strategies to monitor and mitigate the environmental impact associated with increased computational demands and advancing the principle of energy proportionality in AI applications[16].

11 Technopolis Group, Alan Turing Institute. 2022. Review of Digital Research Infrastructure Requirements for AI. See: https://www.turing.ac.uk/sites/default/files/2022-09/ukri-requirements-report_final_edits.pdf (accessed February 6 2024)

12 UKRI. Transforming our world with AI. See: https://www.ukri.org/publications/transforming-our-world-with-ai/ (accessed 6 February 2024)

13 United Nations. 2023 Interim Report: Governing AI for Humanity. See: https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf (accessible 6 February 2024)

14 Lannelongue, L, *et al*. 2023. Greener principles for environmentally sustainable computational science. Nat Comput Sci3, 514–521. (https://doi.org/10.1038/s43588-023-00461-y)

15 The Royal Society. 2023 Privacy Enhancing Technologies. See https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/ (accessed 21 December 2023).

16 The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/ (accessed 21 December 2023).

AREA FOR ACTION: ENHANCE ACCESS TO ESSENTIAL AI INFRASTRUCTURES AND TOOLS

RECOMMENDATION 2

Funders and AI developers should prioritise accessibility and usability of AI tools developed for scientific research

Access to AI does not guarantee its meaningful and responsible use. Complex and high-performance AI tools and methods can be challenging for researchers from non-AI backgrounds to adopt and utilise effectively[17]. Similarly, new skills are needed across the AI lifecycle, such as data scientists who understand the importance of metadata and data curation, or engineers who are familiar with GPU programming for image-based processing.

Taking steps to improve the usability of AI-based tools (eg software applications, libraries, APIs, or general AI systems) should therefore involve a combination of mechanisms that make AI understandable for non-AI experts and build their capacity to use AI responsibly. For example, training should ensure that every scientist is able to recognise when they require specialised data or programming expertise in their teams, or when the use of complex and opaque AI techniques could undermine the integrity and quality of results.

Improving usability can also enhance the role of non-AI scientists as co-designers[18] – as opposed to passive users – who can ensure AI tools meet the needs of the scientific community. Creating conditions for co-design requires bridging disciplinary siloes between AI and domain experts through the development of shared languages, modes of working, and tools.

17 Cartwright H. 2023 Interpretability: Should – and can – we understand the reasoning of machine-learning systems? In: OECD (ed.) *Artificial Intelligence in Science*. OECD. (https://doi.org/10.1787/a8d820bd-en)

18 UKRI. Trustworthy Autonomous Systems Hub. Developing machine learning models with codesign: how everyone can shape the future of AI. See: https://tas.ac.uk/developing-machine-learning-models-with-codesign-how-everyone-can-shape-the-future-of-ai/ (accessed 7 March 2023)

Actions to enhance the usability of AI tools may include:

1. Research institutions and training centres establishing AI literacy curriculums across scientific fields to build researchers' capacity to understand the opportunities, limitations, and adequacy of AI-based tools within their fields and research contexts.
2. Research institutions and training centres establishing comprehensive data literacy curriculums tailored to the specific needs of AI applications in scientific research. This involves building capacity for data management, curation, and stewardship, as well as implementation of data principles such as FAIR (Findable, Accessible, Interoperable, and Reusable) and CARE (Collective benefit, Authority to control, Responsibility, and Ethics)[19].
3. Research funders and AI developers investing in strategies that improve understanding and usability of AI for non-AI experts, with a focus on complex and opaque models[20]. This can include further research on domain-specific explainable AI (XAI) or accessible AI tools that enhance access in resource-constrained research environments[21].
4. Research institutions, research funders, and scientific journals implementing mechanisms to facilitate knowledge translation across domains and meaningful collaboration across disciplines. This requires a combination of cross-discipline training, mentorship, publication outlets and funding (eg through bodies such as the UKRI's Cross-Council Remit Agreement that governs interdisciplinary research proposals)[22].

19 Global Indigenous Data Alliance. Care Principles for Indigenous Data Governance. See https://www.gida-global.org/care (accessed 21 December 2023)

20 Szymanski M, Verbert K, Vanden Abeele V. 2022. Designing and evaluating explainable AI for non-AI experts: challenges and opportunities. In Proceedings of the 16th ACM Conference on Recommender Systems (https://doi.org/10.1145/3523227.3547427)

21 Korot E *et al.* 2021 Code-free deep learning for multi-modality medical image classification. *Nat Mach Intell*. 3, 288–298. (https://doi.org/10.1038/s42256-021-00305-2)

22 UKRI. Get Support For Your Project: If your research spans different disciplines. See: https://www.ukri.org/apply-for-funding/how-to-apply/preparing-to-make-a-funding-application/if-your-research-spans-different-disciplines/ (accessed 13 December 2023)

AREA FOR ACTION: BUILD TRUST IN THE INTEGRITY AND QUALITY OF AI-BASED SCIENTIFIC OUTPUTS

RECOMMENDATION 3

Research funders and scientific communities should ensure that AI-based research meets open science principles and practices to facilitate AI's benefits in science.

A growing body of irreproducible AI and machine learning (ML)-based studies are raising concerns regarding the soundness of AI-based discoveries[23,24]. However, scientists are facing challenges to improve the reproducibility of their AI-based work. These include insufficient documentation released around methods, code, data, or computational environments[25]; limited access to computing to validate complex ML models[26]; and limited rewards for the implementation of open science practices[27]. This poses risks not only to science, but also to society, if the deployment of unreliable or untrustworthy AI-based outputs leads to harmful outcomes[28].

To address these challenges, AI in science can benefit from following open science principles and practices. For example, the UNESCO Recommendation on Open Science[29] offers relevant guidelines to improve scientific rigour, while noting that there is not a one-size-fits-all approach to practising openness across sectors and regions. This aligns well with the growing tendency towards adopting 'gradual' open models that pair the open release of models and data with the implementation of detailed guidance and guardrails to credible risks[30].

Open science principles can also contribute towards more equitable access to the benefits of AI and to building the capacity of a broader range of experts to contribute to its applications for science. This includes underrepresented and under-resourced scholars, data owners, or non-scientist publics.

23 Haibe-Kains B *et al.* 2020 Transparency and reproducibility in artificial intelligence. *Nature.* 586, E14–E16. (https://doi.org/10.1038/s41586-020-2766-y)

24 Kapoor S and Narayanan A. 2023 Leakage and the reproducibility crisis in machine-learning-based science. *Patterns.* 4(9) (https://doi.org/10.1016/j.patter.2023.100804)

25 Pineau, J, *et al*. 2021. Improving reproducibility in machine learning research (a report from the Neurips 2019 Reproducibility program)." Journal of Machine Learning Research 22.164.

26 Bommasani *et al*. 2021. On the opportunities and risks of foundation models. See: https://crfm.stanford.edu/assets/report.pdf (accessed 21 March 2024)

27 UK Parliament, Reproducibility and Research Integrity – Report Summary See: https://publications.parliament.uk/pa/cm5803/cmselect/cmsctech/101/summary.html (accessed 7 February 2024)

28 Sambasivan, N, *et al* 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

29 UNESCO Recommendation on Open Science. 2021. See: https://www.unesco.org/en/legal-affairs/recommendation-open-science (accessed 6 February 2024)

30 Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (https://doi.org/10.48550/arXiv.2302.04844)

Further work is needed to understand the interactions between open science and AI for science, as well as how to minimise safety and security risks stemming from the open release of models and data.

Actions to promote the adoption of open science in AI-based science may include:

1. Research funders and research institutions incentivising the adoption of open science principles and practices to improve reproducibility of AI-based research. For example, by allocating funds to open science and AI training, requesting the use of reproducibility checklists[31] and data sharing protocols as part of grant applications, or by supporting the development of community and field-specific reproducibility standards (eg TRIPOD-AI[32]).
2. Research institutions and journals rewarding and recognising open science practices in career progression opportunities. For example, by promoting the dissemination of failed results, accepting pre-registration and registered reports as outputs, or recognising the release of datasets and documentation as relevant publications for career progression.
3. Research funders, research institutions and industry actors incentivising international collaboration by investing in open science infrastructures, tools, and practices. For example, by investing in open repositories that enable the sharing of datasets, software versions, and workflows, or by supporting the development of context-aware documentation that enables the local adaptation of AI models across research environments. The latter may also contribute towards the inclusion of underrepresented research communities and scientists working in low-resource contexts.
4. Relevant policy makers considering ways of deterring the development of closed ecosystems for AI in science by, for example, mandating the responsible release of benchmarks, training data, and methodologies used in research led by industry.

31 McGill School of Computer Science. The Machine Learning Reproducibility Checklist v2.0. See: https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf (accessed 21 December 2023).

32 Collins G *et al*. 2021 Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ open*, 11(7), e048008. (https://doi.org/10.1136/bmjopen-2020-048008)

AREA FOR ACTION: ENSURE SAFE AND ETHICAL USE OF AI IN SCIENTIFIC RESEARCH

RECOMMENDATION 4

Scientific communities should build the capacity to oversee AI systems used in science and ensure their ethical use for the public good

The application of AI across scientific domains requires careful consideration of potential risks and misuse cases. These can include the impact of data bias[33], data poisoning[34], the spread of scientific misinformation[35,36], and the malicious repurposing of AI models[37]. In addition to this, the resource-intensive nature of AI (eg in terms of energy, data, and human labour) raises ethical questions regarding the extent to which AI used by scientists can inadvertently contribute to environmental and societal harms.

Ethical concerns are compounded by the uncertainty surrounding AI risks. As of late 2023, public debates regarding AI safety had not conclusively defined the role of scientists in monitoring and mitigating risks within their respective fields. Furthermore, varying levels of technical AI expertise among domain experts, and the lack of standardised methods for conducting ethics impact assessments, limit the ability of scientists to provide effective oversight[38]. Other factors include the limited transparency of commercial models, the opaque nature of ML-systems, and how the misuse of open science practices could heighten safety and security risks[39,40].

As AI is further integrated into science, AI assurance mechanisms[41] are needed to maintain public trust in AI and ensure responsible scientific advancement that benefits humanity. Collaboration between AI experts, domain experts and researchers from humanities and science, technology, engineering, the arts, and mathematics (STEAM) disciplines can improve scientists' ability to oversee AI systems and anticipate harms[42].

33 Arora, A, Barrett, M, Lee, E, Oborn, E and Prince, K 2023 Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. *Information and Organization*, **33**. (https://doi.org/10.1016/j.infoandorg.2023.100478)

34 Verde, L., Marulli, F. and Marrone, S., 2021. Exploring the impact of data poisoning attacks on machine learning model reliability. *Procedia Computer Science*, **192**. 2624-2632. (https://doi.org/10.1016/j.procs.2021.09.032)

35 Truhn D, Reis-Filho J.S. & Kather J.N. 2023 Large language models should be used as scientific reasoning engines, not knowledge databases. *Nat Med* **29**, 2983–2984. (https://doi.org/10.1038/s41591-023-02594-z)

36 The Royal Society. 2024. Red teaming large language models (LLMs) for resilience to scientific disinformation. See https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/

37 Kazim, E and Koshiyama, A.S 2021 A high-level overview of AI ethics. *Patterns*, **2**. (https://doi.org/ 10.1016/j.patter.2021.100314)

38 Wang H *et al*. 2023 Scientific discovery in the age of artificial intelligence. *Nature*, **620**. 47-60. (https://doi.org/10.1038/s41586-023-06221-2)

39 Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (https://doi.org/10.48550/arXiv.2302.04844)

40 Vincent J. 2023 OpenAI co-founder on company's past approach to openly sharing research: 'We were wrong'. *The Verge*. See https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview (accessed 21 December 2023).

41 Brennan, J. 2023. AI assurance? Assessing and mitigating risks across the AI lifecycle. Ada Lovelace Institute. See https://www.adalovelaceinstitute.org/report/risks-ai-systems/ (accessed 30 September 2023)

42 The Royal Society. 2023. Science in the metaverse: policy implications of immersive technology. See https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/

Similarly, engaging with communities represented in or absent from AI training datasets, can improve the current understanding of possible risks and harms behind AI-based research projects.

Actions to support the ethical application of AI in science can include:

1. Research funders and institutions investing in work that operationalises and establishes domain-specific taxonomies[43] of AI risks in science, particularly sensitive fields (eg chemical and biological research).
2. Research funders, research institutions, industry actors, and relevant scientific communities embracing widely available ethical frameworks for AI, as reflected in the UNESCO Recommendation on the Ethics of Artificial Intelligence[44], or the OECD's Ethical Guidelines for Artificial Intelligence[45], and implementing practices that blend open science with safeguards against potential risks.
3. Funders, research institutions and training centres providing AI ethics training and building the capacity of scientists to conduct foresight activities (eg horizon scanning), pre-deployment testing (eg red teaming), or ethical impact assessments of AI models to identify relevant risks and guardrails associated with their field.
4. Research funders, research institutions, and training centres supporting the development of interdisciplinary and participatory approaches to safety auditing, ensuring the involvement of AI and non-AI scientists, and affected communities in the evaluation of AI applications for scientific research.

43 Weidinger L, *et al*. 2022 Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214-229. (https://doi.org/10.1145/3531146.3533088)

44 UNESCO. 2022. Recommendation on the ethics of artificial intelligence. See: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics (accessed 5 March 2024)

45 OECD. Ethical guidelines for artificial intelligence. See: https://oecd.ai/en/catalogue/tools/ethical-guidelines-for-artificial-intelligence (accessed 5 March 2024)

Introduction

Scope of the report

Science in the age of AI explores how AI is transforming the nature and methods of scientific research. It focuses on the impact of deep learning methods and generative AI applications and explores cross-cutting considerations around research integrity, skills, and ethics. While AI is transforming a wide range of fields – including the social sciences and humanities – this report provides examples focused on physical and biological sciences.

The report addresses following questions:

- How are AI-driven technologies transforming the methods and nature of scientific research?
- What are the opportunities, limitations, and risks of these technologies for scientific research?
- How can relevant stakeholders (governments, universities, industry, research funders, etc) best support the development, adoption, and uses of AI-driven technologies in scientific research?

Each chapter provides evidence gathered from interviews, roundtables, workshops, and commissioned research undertaken for this report to answer these questions. The findings are presented as follows:

- **Chapter 1** provides a descriptive review of how recent developments in AI (in Machine Learning (ML), deep neural networks, and natural language processing in particular) are changing methods, processes, and practices in scientific research.
- **Chapter 2** details key challenges for research integrity in AI-based research. It tackles issues around transparency of AI models and datasets, explainability and interpretability, and barriers to verifying the reproducibility of results.
- **Chapter 3** addresses interdisciplinary collaboration and emerging research skills in AI-driven scientific research. It examines challenges such as siloed academic cultures and data infrastructures, explores opportunities for collaboration across fields, and characterises the need for data, AI, and ethics upskilling and training programmes.
- **Chapter 4** addresses the growing role of the private sector in AI-based research. It considers the opportunities and challenges related to the private sector's impact on the public sector, as well as examples of cross-sector collaboration.
- **Chapter 5** addresses research ethics and safety in AI-based research. It highlights the need for oversight to prevent downstream harms related to ethical challenges, safety, and security risks. Examples include data bias, hallucinations, or the repurposing of datasets and models with malicious intent.

The report also includes case studies on the application of AI for climate science, material science, and rare disease diagnosis. These cases also explore challenges related to researchers' access to data (Case study 1), the implications of lab automation (Case study 2), and the emerging research ethics considerations in applications of AI in science (Case study 3).

Target audiences

The following audiences should find this report useful:

- Scientists and research funders navigating the changing role of AI in science. This report offers an overview of the opportunities and challenges associated with the integration of increasingly complex techniques in scientific research.
- AI experts and developers. This report presents a case for further interdisciplinary collaboration with scientific domain experts and for strengthening cross-sector collaboration.
- Policy makers and regulators involved in shaping AI and data strategies. Evidence gathered in this report can contribute to strategies that promote responsible AI development, address ethical concerns, and support scientific progress.
- General public seeking to understand future directions and applications of AI. This report contributes towards informing the public regarding opportunities and challenges associated with AI adoption, as well as the broader societal implications of advancing this technology.

Readers need not have a technical background on AI to read this report.

Glossary of key terms

This report draws on concepts from data science and AI fields[46]. Included here is an overview of key terms used throughout.

Artificial intelligence (AI): The development and study of machines capable of performing tasks that conventionally required human cognitive abilities. It encompasses various aspects of intelligence, such as reasoning, decision-making, learning, communication, problem-solving, and physical movement. AI finds widespread application in everyday technology such as virtual assistants, search engines, navigation systems, and online banking.

Artificial neural networks (ANNs): Artificial intelligence systems inspired by the structure of biological brains, consisting of interconnected computational units (neurons) capable of passing data between layers. Today, they excel in tasks such as face and voice recognition, with multiple layers contributing to problem-solving capabilities. See also 'deep learning'.

Deep learning (DL): A form of machine learning utilising computational structures known as 'artificial neural networks' to automatically recognise patterns in data and produce relevant outputs. Inspired by biological brains, deep learning model are proficient at complex tasks such as image and speech recognition, powering applications like voice assistants and autonomous vehicles. See 'Artificial neural networks'.

Foundation model: A machine learning model trained on extensive data, adaptable for diverse applications. Common examples include large language models, serving as the basis for various AI applications like chatbots.

46 The Alan Turing Institute. Defining data science and AI. See: https://www.turing.ac.uk/news/data-science-and-ai-glossary (accessed 1 March 2024)

Generative AI: AI systems generating new text, images, audio, or video in response to user input using machine learning techniques. These systems, often employing Generative adversarial networks (GANs), create outputs closely resembling human-created media, resulting in outputs that are often indistinguishable from human-created media. See 'Generative adversarial networks'.

General adversarial networks (GANs): A machine learning technique that produces realistic synthetic data, like deepfake images, indistinguishable from its training data. It consists of a generator and a discriminator. The generator creates fake data, while the discriminator evaluates it against real data, helping the generator improve until the discriminator can't differentiate between real and fake.

Human-in-the-loop (HITL): A hybrid system comprising of human and artificial intelligence that allows for human intervention, such as training or fine-tuning the algorithm, to enhance the systems output. Combining the strengths of both human judgment and machine capabilities can make up for the limitations of both.

Large language models (LLM): Foundation models trained on extensive textual data to perform language-related tasks, including chatbots and text generation. They are part of a broader field of research called natural language processing, and are typically much simpler in design than smaller, more traditional language models.

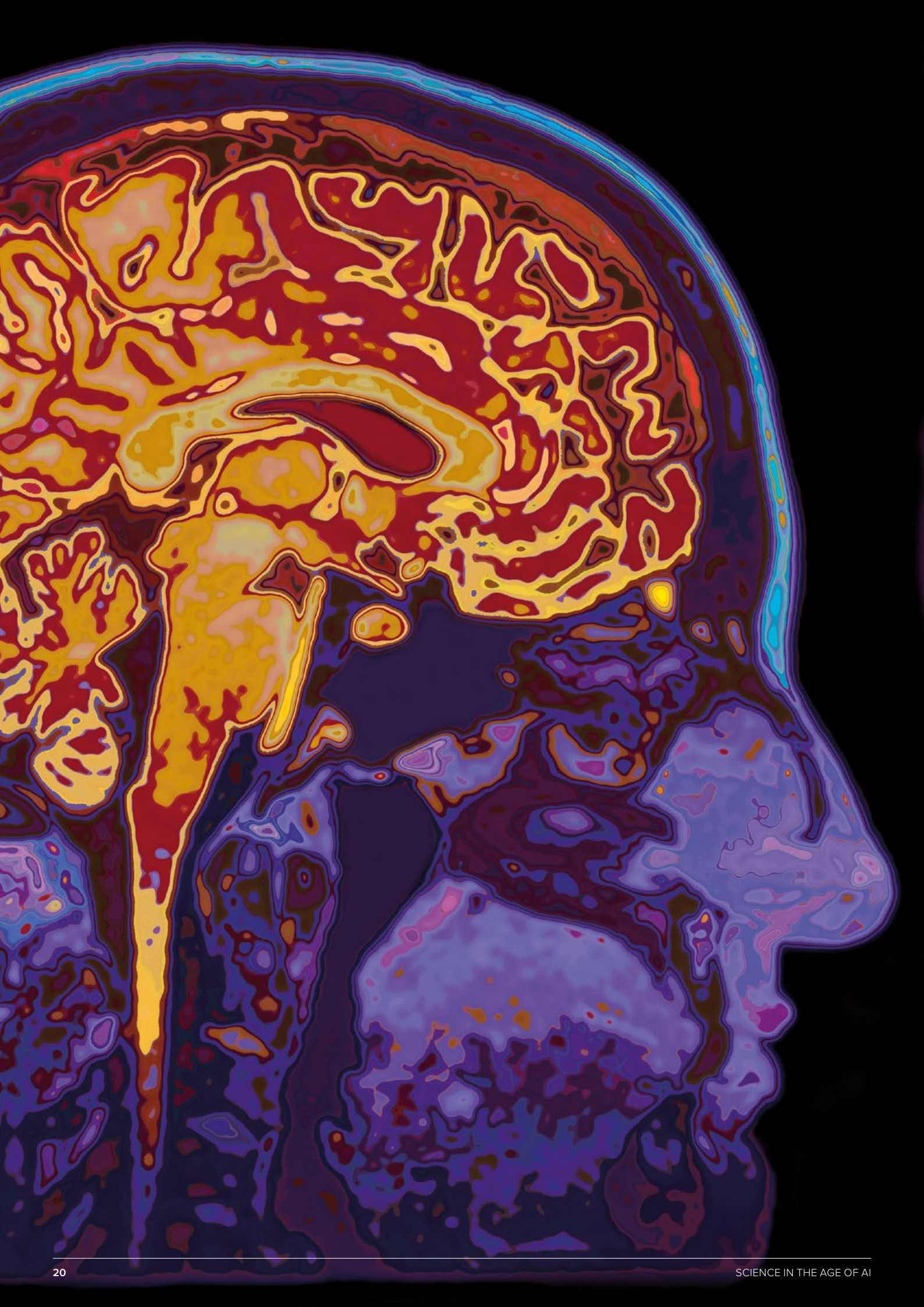
Machine learning (ML): A field of artificial intelligence involving algorithms that learn patterns from data and apply these findings to make predictions or offer useful outputs. It enables tasks like language translation, medical diagnosis, and robotics navigation by analysing sample data to improve performance over time.

Privacy-enhancing technologies (PETs): An umbrella term covering a broad range of technologies and approaches that can help mitigate data security and privacy risks[47].

Synthetic data: Data that is modelled to represent the statistical properties of original data; new data values are created which, taken as a whole, preserve relevant statistical properties of the 'real' dataset[48]. This allows for training models without accessing real-world data.

47 The Royal Society. 2019 Protecting privacy in practice. See https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/ (accessed 1 March 2024).

48 *Ibid.*



Chapter one
How artificial intelligence is transforming scientific research

Left
MRI image. © iStock / MachineHeadz.

How artificial intelligence is transforming scientific research

"We are really reliant on [machine learning] just to make [our experiments] work. It has become embedded into what we're doing. If you took machine learning out of our pipeline, it would fall apart."

Royal Society roundtable participant

The large-scale data analysis and pattern recognition capabilities of artificial intelligence (AI) present significant opportunities for advancing scientific research and discovery. New developments in machine learning, in particular, are enabling researchers to map deforestation down to an individual tree[43]; pharmaceutical companies to develop new therapies[44]; and technology companies to discover new materials[45]. These developments present novel opportunities and challenges to the nature and method of scientific investigation.

Drawing on insights from roundtables, interviews, and commissioned research, this chapter outlines how AI is changing the scientific endeavour.

AI in science: an overview

A commissioned analysis of the use of AI in science (based on published academic literature and focused on breadth, rather than depth of techniques) shows that applications of AI are found across all STEM fields[46]. There is a concentration in certain fields such as medicine; materials science; robotics; genetics; and, unsurprisingly, computer science. The physical sciences and medicine appear to dominate when it comes to the use of AI-related technologies. The most prominent AI techniques across STEM fields include artificial neural networks (ANNs); machine learning (ML) (including deep learning (DL)); natural language processing; and image recognition.

In the physical sciences (eg, physics, chemistry, astronomy), AI is being applied as a method for extracting information from rapidly accumulating data streams, (eg data generated at the Large Hadron Collider[47]); to identify patterns in very large datasets; and for modelling physical experiments. In the health sciences (eg, medicine, dentistry, veterinary sciences), it is being employed as a technique to aid disease detection and prediction; to support clinical decision-making; and to enhance surgery, training, and robotics. In the life sciences, it is being used to analyse data from sensors; to support crop and water management; and to predict the 3D structures of proteins.

AI and methods of scientific research

Recent developments in AI suggest there may be transformational changes to the methods of scientific research. These changes centre on making existing tasks more efficient, altering processes to generate knowledge, or enabling new mechanisms of discovery.

The following examples emerged in the research activities undertaken for this report.

43 Eyres A *et al*. 2024 LIFE: A metric for quantitatively mapping the impact of land-cover change on global extinctions. Cambridge Open Engage. (https://doi.org/10.33774/coe-2023-gpn4p-v4).

44 Paul, D, Sanap, G, Shenoy, S, Kalyane, D, Kalia, K, Tekade, R. K. 2021 Artificial intelligence in drug discovery and development. *Drug discovery today*, **26**. 80–93. (https://doi.org/10.1016/j.drudis.2020.10.010)

45 Merchant, A, Batzner, S, Schoenholz, SS, Aykol, M, Cheon, G, Cubuk, ED. 2023 Scaling deep learning for materials discovery. *Nature*, **624**. 80–85. (https://doi.org/10.1038/s41586-023-06735-9)

46 Berman B, Chubb J, and Williams K, 2024. The use of artificial intelligence in science, technology, engineering, and medicine. The Royal Society. https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/ (accessed 7 May 2024)

47 CERN. The Large Hadron Collider. See https://home.cern/science/accelerators/large-hadron-collider (accessed 22 April 2024)

1. **Growing use of deep learning across fields**
   The application of deep learning (DL) is transforming data analysis and knowledge generation. Its use to automatically extract and learn features from raw data, process extensive datasets and recognise patterns efficiently outperforms linear ML-based models[48]. DL has found applications in diverse fields including healthcare, aiding in disease detection and drug discovery, or climate science, assisting in modelling climate patterns and weather detection. A landmark example is the application of DL by Google DeepMind to develop AlphaFold, a protein-folding prediction system that solved a 50-year-old challenge in biology decades earlier than anticipated[49].

   Developing accurate and useful DL-based models is challenging due to its black-box nature and variations in real-world problems and data. This limits their explanatory power and reliability as scientific tools[50]. (See Chapter 2).

2. **Obtaining insights from unstructured data**
   A major challenge for researchers is utilising unstructured data (data that does not follow a specific format or structure, making it more challenging to process, manage and use to find patterns). The ability to handle unstructured data makes DL effective for tasks that involve image recognition and natural language processing (NLP).

   In healthcare, for example, data can be detailed; multi-modal and fragmented[51]. It can include images, text assessments, or numerical values from assessments and readings. Data collectors across the healthcare system may record this data in different formats or with different software. Bringing this data together, and making sense of it, can help researchers make predictions and model potential health interventions. Similarly, generative AI models can contribute towards generating and converting data into different modes and standards, that are not limited to the type of data fed into the algorithm[52].

> "We have the capacity to record much more [data] than before. We live in a data deluge. So, the hope is that machine learning methods will help us make sense of that, and then drive genuine, scientific hypotheses."
>
> **Royal Society roundtable participant**

48 Choudhary, A, Fox, G, Hey, T. 2023. Artificial intelligence for science: A deep learning revolution. *World Scientific Publishing Co. Pte Ltd*. (https://doi.org/10.1142/13123)

49 Google DeepMind. AlphaFold. See: https://deepmind.google/technologies/alphafold/ (accessed 5 March 2024)

50 Sarker, I. H. 2021. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, **2**. 420. (https://doi.org/10.1007/s42979-021-00815-1)

51 Healy, M. J. R. 1973. What Computers Can and Cannot Do. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *184*(1077), 375–378. (https://doi.org/10.1098/rspb.1973.0056)

52 World Health Organization. 2024. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. See: https://www.who.int/publications/i/item/9789240084759 (accessed 5 March 2024)

Other techniques such as causal machine learning (methods to estimate cause and effect in data)[53] can help process unstructured data by learning complex nonlinear relations between variables[54,55]. Platforms claiming to be able to gain insights from unstructured data include Benevolent AI[56] (using unstructured data from biomedical literature) and Microsoft's Project Alexandria[57] (focused mainly on enterprise knowledge).

3. Large-scale, multi-faceted simulations

The generative capability of AI tools to learn from existing content and generate predictions of new content, provides scientists with the opportunity to run accurate predictions. The generation of simulations and synthetic data[58] (artificially generated data) or digital twins[59] (virtual representations of physical assets) are examples of how AI-based tools can be used to train systems[60]. For molecular research, this involves using deep neural networks that use data about how molecules interact to accurately simulate the behaviour at the atomic level [61]. The use of synthetic data, especially privacy-preserving synthetic data, can also help mitigate the challenge of data bias and protect individuals' privacy[62].

4. Expediting information synthesis

Large language models (LLMs) and NLP techniques are increasingly being used to accelerate text-based tasks such as academic writing[63], conducting literature reviews, or producing summaries[64].

53 Kaddour J, Lynch A, Liu Q, Kusner M J, Silva R. 2022. Causal machine learning: A survey and open problems. *arXiv preprint* (https://doi.org/10.48550/arXiv.2206.15475)

54 Sanchez P, Voisey J, Xia T, Watson H, O'Neil A, and Tsaftaris, S. 2022 Causal machine learning for healthcare and precision medicine. *R. Soc open sci.* **9**: 220638 (https://doi.org/10.1098/rsos.220638)

55 Royal Society roundtable on large language models, July 2023.

56 Benevolent AI. 2019 Extracting existing facts without requiring any training data or hand-crafted rules. See https://www.benevolent.com/news-and-media/blog-and-videos/extracting-existing-facts-without-requiring-any-training-data-or-hand-crafted-rules/ (accessed 21 December 2023).

57 Rajput S, Winn J, Moneypenny N, Zaykov Y, and Tan C. 2021 Alexandria in Microsoft Viva Topics: from big data to big knowledge. 26 April 2021. See https://www.microsoft.com/en-us/research/blog/alexandria-in-microsoft-viva-topics-from-big-data-to-big-knowledge/ (accessed 21 December 2023).

58 Jordon *et al*. 2023 Synthetic Data – what, why and how? See https://royalsociety.org/news-resources/projects/privacy-enhancing-technologies/ (accessed 21 December 2023)

59 The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/ (accessed 21 December 2023).

60 Jordon *et al*. 2023 Synthetic Data – what, why and how? See https://royalsociety.org/news-resources/projects/privacy-enhancing-technologies/ (accessed 21 December 2023).

61 Zhang L, Han J, Wang H, Car R, Weinan E. 2018 Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. Phys Rev Lett. 2018 Apr 6;120(14):143001. (https://doi.org/10.1103/PhysRevLett.120.143001. PMID: 29694129)

62 The Royal Society. 2023 From privacy to partnership. See https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/ (accessed 21 December 2023).

63 Lin Z. 2023 Why and how to embrace AI such as ChatGPT in your academic life. *R. Soc. Open Sci.***10:** 230658 230658 (https://doi.org/10.1098/rsos.230658)

64 Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 2023 Feb 19;15(2):e35179. (https://doi: 10.7759/cureus.35179)

Examples of automated literature review tools include Semantic Scholar[65], Elicit[66], and Consensus[67]. It is also available on prominent platforms such as GPT4 and Gemini. Beneficial use cases include using LLMs to improve the quality of academic writing, assist with translation, or emulate specific writing styles (eg producing lay summaries). Beyond academic texts, they can also be used to streamline administrative tasks and assist in drafting grant applications. These tools could also improve accessibility for researchers from diverse backgrounds (eg non-English speakers and neurodivergent individuals) who consume and produce academic content in multiple languages and formats[68].

These tools also have limitations including the potential to exacerbate biases from the training data (eg bias towards positive results[69], language biases[70] or geographic bias[71]), inaccuracies and unreliable scientific inputs[72]. As a writing tool they also have a limited ability to grasp nuanced value judgments, assist in scientific meaning-making[73], or articulate the complexities of scientific research[74]. There are also concerns that the use of LLMs for academic writing risks diminishing creative and interdisciplinary aspects of scientific discovery[75]. Additionally, there are questions around the impact of LLMs on intellectual property (IP).

5. **Addressing complex coding challenges**
   Developing computational analysis software code has become an important aspect of the modern scientific endeavour. For example, LLMs – which are designed to analyse text inputs and generate responses that they determine are likely to be accurate – can be used for generating software code in various coding languages. This presents an opportunity for scientific researchers to convert code from one computer language to another, or between applications[76].

65 Semantic Scholar: AI-powered research tool. See https://www.semanticscholar.org/ (accessed 21 December 2023).

66 Elicit: The AI research assistant. See https://elicit.com/ (accessed 21 December 2023).

67 Consensus: AI search engine for research. See https://consensus.app/ (accessed 21 December 2023).

68 Royal Society and Department for Science, Innovation, and Technology workshop on horizon scanning AI safety risks across scientific disciplines, 2023.

69 Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See https://royalsociety.org/current-topics/ai-data/. (accessed 7 May 2024).

70 Barrot JS, 2023. Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, **57**.100745.

71 Skopec M, Issa H, Reed J, Harris M. 2020. The role of geographic bias in knowledge diffusion: a systematic review and narrative synthesis. *Research integrity and peer review*, **5**. 1-14. (https://doi.org/10.1186/s41073-019-0088-0.)

72 Sanderson K. 2023. GPT-4 is here: what scientists think. Nature, 615.773. 30 March 2023. See https://www.nature.com/articles/d41586-023-00816-5.pdf (accessed 21 December 2023)

73 Birhane A, Kasirzadeh A, Leslie D, Wachter S. 2023. Science in the age of large language models. Nature Reviews Physics, 1-4 (https://doi.org/10.1038/s42254-023-00581-4)

74 Bender E, Koller A. 2020 Climbing towards NLU: on meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 5185–5198

75 Royal Society and Department for Science, Innovation, and Technology workshop on horizon scanning AI safety risks across scientific disciplines, 2023.

76 Royal Society roundtable on large language models, July 2023.

Even if the output is not accurate on a first attempt, these models can be used as coding assistants to help identify coding mistakes, make suggestions, and save time. Prominent examples include Microsoft's Copilot[77]; OpenAI's GPT4[78]; Meta's Code Llama[79]; and Google DeepMind's Gemini.[80]

6. Task automation

AI tools can automate a range of time and labour-intensive tasks within the scientific workflow.[81] Automation can lead to productivity gains for scientists[82] and unlock the potential to test diverse hypotheses beyond human capability. For example, in 2023, Google DeepMind claimed two such examples: FunSearch[83], and GNoME[84].

The use of robotic research assistants is also contributing to the automation of laboratory workflows (See Case Study 2). In 2009, a robot developed by Aberystwyth University became the first machine to independently discover new scientific knowledge[85]. The robot was programmed to independently design experiments, record and evaluate results, and develop new questions – automating the entire research workflow[86]. Building on this breakthrough, 'robot scientists' continue to be developed to speed up the discovery process, while reducing costs, uncertainty, and human error in labs[87].

As research becomes more automated, there are concerns that future generations of scientists may become de-skilled in core skills such as hypothesis generation, experimental design, and contextual interpretation[88]. Methodological transparency and understanding of cause-effect relationships could also decline, and an overemphasis on computational techniques risks disengaging scientists who seek creative outlets in their work[89].

77 GitHub. Copilot – Your AI pair programmer. See https://github.com/features/copilot (accessed 21 December 2023).

78 Open AI. GPT4. See https://openai.com/gpt-4 (accessed 21 December 2023).

79 Meta. 2023 Introducing Code Llama, a state-of-the-art large language model for coding. *Meta*. 24 August 2023. See https://ai.meta.com/blog/code-llama-large-language-model-coding/ (accessed 21 December 2023).

80 Google DeepMind. Gemini. See https://deepmind.google/technologies/gemini/#introduction (accessed 21 December 2023).

81 Xie, Y, Sattari, K, Zhang, C, & Lin, J. 2023 Toward autonomous laboratories: Convergence of artificial intelligence and experimental automation. *Progress in Materials Science,* **132**. 101043. (https://doi.org/10.1016/j.pmatsci.2022.101043)

82 OECD. 2023. Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*, OECD Publishing, Paris* (https://doi.org/10.1787/a8d820bd-en).

83 Fawzi A and Paredes B. 2023. FunSearch: Making new discoveries in mathematical sciences using Large Language Models. *Google DeepMind*. See https://deepmind.google/discover/blog/funsearch-making-new-discoveries-in-mathematical-sciences-using-large-language-models/ (accessed 21 December 2023).

84 Merchant A and Cubuk E. 2023 Millions of new materials discovered with deep learning. *Google DeepMind*. See https://deepmind.google/discover/blog/millions-of-new-materials-discovered-with-deep-learning/ (accessed 21 December 2023).

85 University of Cambridge. Robot scientist becomes first machine to discover new scientific knowledge. See: https://www.cam.ac.uk/research/news/robot-scientist-becomes-first-machine-to-discover-new-scientific-knowledge (accessed 3 March 2024)

86 Sparkes A *et al*. 2010. Towards Robot Scientists for autonomous scientific discovery. Autom Exp 2, 1 (https://doi.org/10.1186/1759-4499-2-1)

87 University of Cambridge. Artificially-intelligent Robot Scientist 'Eve' could boost search for new drugs. See: https://www.cam.ac.uk/research/news/artificially-intelligent-robot-scientist-eve-could-boost-search-for-new-drugs (accessed 7 March 2024)

88 Lin Z. 2023 Why and how to embrace AI such as ChatGPT in your academic life. R Soc Open Sci. 2023 Aug 23;10(8):230658 (https://doi.org/ 10.1098/rsos.230658.)

89 Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See https://royalsociety.org/current-topics/ai-data/. (accessed 7 May 2024)

AI and the nature of scientific research

Beyond the impact of AI on the methods of scientific research, there is a potentially transformative impact on the nature of the scientific endeavour itself. These impacts primarily relate to the prevalence of big data-led research, reliance on computing power and new ways of organising skills and labour in the scientific process.

Drawing on the activities undertaken for this report, the following six themes emerged as key impacts of AI on the nature of scientific research.

1. **Computers and labour as foundational AI infrastructures**
   An assemblage of digital infrastructure and human labour underpins major AI applications[90]. The digital infrastructure refers to devices which collect data, personal computers which they are analysed on, and supercomputers which power large-scale data analysis. The human labour refers to the act of data collection, cleaning, and labelling, as well as the act of design, testing, and implementation. The types of digital infrastructure required includes supercomputers (eg those included in HPC-UK[91] and the EuroHPC JU[92]); privacy enhancing technologies;[93] and data storage facilities (eg data centres). Cloud-based solutions, which do not require users to own physical infrastructure (eg to store data) include Amazon Web Services[94] and Oracle Cloud Infrastructure.[95]

2. **Domination of big data centric research**
   The ability to collect big data (large and heterogeneous forms of data that have been collected without strict experimental design[96]) and combine these with other datasets has presented clear and significant opportunities for the scientific endeavour. The value being gained from applying AI to these datasets has already provided countless examples of positive applications from mitigating the impact of COVID-19 to combating climate change (See Case Study 3)[97]. This is likely to continue to reshape the research endeavour to be more AI and big data-centric[98]. The ability to engage in data-centric research, however, remains dependent on access to computing infrastructure that enables processing of large heterogenous datasets.

   The domination of big data centric research also has implications for research in which only incomplete or small data is available. Without careful governance, it risks reducing research investment and support in priority areas (eg subjects or regions) where primary data collection at that scale is limited, difficult

90 Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/ (accessed 7 May 2024)

91 HPC-UK. UK HPC Facilities. See https://www.hpc-uk.ac.uk/facilities/ (accessed 21 December 2023).

92 The European High Performance Computing Joint Undertaking. See https://eurohpc-ju.europa.eu/index_en (accessed 21 December 2023).

93 The Royal Society. Privacy Enhancing Technologies. See https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/ (accessed 21 December 2023).

94 Amazon Web Services. Cloud Computing Services. See https://aws.amazon.com/ (accessed 21 December 2023).

95 Oracle. Cloud Infrastructure. See https://www.oracle.com/cloud/ (accessed 21 December 2023).

96 The Royal Society. 2017 Machine Learning: The power and promise of computers that learn by example. See https://royalsociety.org/topics-policy/projects/machine-learning/ (accessed 21 December 2023).

97 The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/ (accessed 21 December 2023).

98 Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See https://royalsociety.org/current-topics/ai-data/. (accessed 7 May 2024)

or not desirable. It is also likely to increase attention on techniques such as data augmentation and the use of synthetic data. The case of rare disease research (See Case Study 1) illustrates applications of AI in small data research.

3. Open vs closed science

Open science, which seeks to open the entire research and publication process (including but not limited to open data; open protocols; open code; and transparent peer review), is a principle and practice advocated for by the Royal Society, and others[99]. It is also promoted by major technology companies including Meta and OpenAI, although this has been challenged as 'aspirational' or, even, 'marketing' rather than a technical descriptor[100]. As well as providing transparency, open science approaches can enable replication of experiments, wider public scrutiny of research products[101] and further the right of everyone to share in scientific advancement[102].

However, the increasing use of proprietary AI presents challenges for open science. Researchers are increasingly relying on tools developed and maintained by private companies (see Chapter 4), even though the inner workings may remain opaque[103]. This is exacerbated by the opacity of training data which underpins prominent AI tools. Poor transparency risks limiting the utility of AI tools for solving real world problems as policymakers and scientists may not consider AI-generated results as reliable for important decisions[104]. It also undermines efforts to detect and scrutinise negative impacts or discriminatory effects[105].

A fully open approach that prompts the release of datasets and models without guardrails or guidance may not be desirable either, as datasets or models can be manipulated by bad actors[106]. Context-specific and AI-compatible open science approaches are needed to boost oversight and transparency[107,108].

99 The Royal Society. Open Science. See https://royalsociety.org/journals/open-access/open-science/ (accessed 21 December 2023).

100 Widder D, West S, Whittaker M. 2023 Open (for business): Big tech, concentrated power, and the political economy of Open AI. *SSRN*. (https://doi.org/10.2139/ssrn.4543807)

101 The Royal Society. 2022 The online information environment. See https://royalsociety.org/topics-policy/projects/online-information-environment/ (accessed 21 December 2023).

102 UNESCO Recommendation on Open Science. 2021. See: https://www.unesco.org/en/legal-affairs/recommendation-open-science (accessed 6 February 2024)

103 Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/ (accessed 7 May 2024)

104 UNESCO Recommendation on Open Science. 2021. See: https://www.unesco.org/en/legal-affairs/recommendation-open-science (accessed 6 February 2024)

105 UNESCO Recommendation on ethics of artificial intelligence. 2022. See: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence (accessed 6 February 2024)

106 Vincent J. 2023 OpenAI co-founder on company's past approach to openly sharing research: 'We were wrong'. The Verge. 15 March 2023. See https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview (accessed 21 December 2023)

107 House of Lords. 2024 Large language models and generative AI. Report of Session 2023-24. See: https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/5402.htm (accessed 2 February 2024)

108 Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (https://doi.org/10.48550/arXiv.2302.04844)

4. Challenging notions of transparency and explainability

The scientific method can be generally described as the act of creating a hypothesis, conducting an experiment, recording its outcome, and refining the original hypothesis according to the results. This concept dates back more than a thousand years ago to Hasan Ibn al-Haytham, who emphasised the need for experimental data and reproducibility of results.[109] Underpinning this, and other approaches to scientific methodology is "the search for explanations as a fundamental aim of science"[110].

This is being challenged by recent developments in AI due to the non-linear relationships which can be derived from big data and the general black-box nature of AI tools[111,112]. These discoveries could be transformational for society. If, however, researchers develop an overreliance[113] on AI for the interpretation and analysis of results, they may be unable to explain how conclusions were reached or provide the information required to reproduce a study.[114] This would not meet the threshold of how science is traditionally accepted to be undertaken as peers would struggle to confirm or falsify a hypothesis. (See Chapter 2 for further details on challenges AI poses for explainability and reproducibility).

5. Science as an interdisciplinary endeavour

Successful application of AI in scientific research, and its translation to real-world value, requires interdisciplinary skills and understanding. Computer scientists who wish to apply AI to solve major scientific problems need to be able to evaluate AI models in other research fields (eg health, climate science). Similarly, non-computer scientists need to understand how to effectively use AI tools and techniques for their experiments. Integrating knowledge from various fields and knowledge systems can also lead to more accurate models and foster curiosity-driven research (beyond commercially-driven interests)[115].

109 Al-Khalili J. 2009 The 'first true scientist'. *BBC News*. 4 January 2009. See http://news.bbc.co.uk/1/hi/sci/tech/7810846.stm (accessed 21 December 2023).

110 Maxwell N. 1972 A Critique of Popper's Views on Scientific Method. *Philosophy of Science, 39*(2), 131-152. (doi:10.1086/288429)

111 Succi S, Coveney P. 2019 Big data: The end of the scientific method? Phil. Trans. R. Soc. A. 377: 20180145. (https://doi.org/10.1098/rsta.2018.0145)

112 The Royal Society. 2019 Explainable AI: the basics. See https://royalsociety.org/topics-policy/projects/explainable-ai/ (accessed 21 December 2023).

113 Buçinca Z, Malaya M, Gajos K. 2021 To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proc. ACM Hum.-Comput. Interact. 5:188. (https://doi.org/10.1145/3449287)

114 Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See https://royalsociety.org/current-topics/ai-data/. (accessed 7 May 2024)

115 The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023.

The UK's eScience initiative (2001 – 2008)[116,117] stands out as an effort to cultivate interdisciplinary collaboration by fostering a culture where scientists and computer science experts work together. Ongoing initiatives like Alan Turing Institute[118] and Arizona State University's School of Sustainability[119] also continue to champion interdisciplinary approaches.

However, interdisciplinarity is stifled by siloed institutions and insufficient career progression opportunities. Interdisciplinarity need not be limited to natural sciences, with value to be gained from scientists working with researchers in the arts, humanities, and social sciences. An example of this includes the importance of artists in the user experience design of immersive environments[120,121] (See chapter 3 for further details on interdisciplinarity in AI-based research).

6. Blending human expertise with AI automation

The turn to automation offers opportunities to combine human expertise with efficiencies enabled by AI. AI can be used to either complement the human scientist by assisting or augmenting human capability; or to develop autonomous mechanisms for discovery (See Figure 1)[122]. Across this spectrum, the human scientist remains essential for contextual scientific understanding. The growing use of AI tools also risks making scientists vulnerable to 'illusions of understanding' in which only a limited set of viewpoints and methods are represented in outputs[123]. There is a need to further understand "human-in-the-loop" approaches that recognise AI as complementary to human judgment and the role of human intervention to ensure the quality of outputs.

116 Hey T, Trefethen A. 2002 The UK e-Science Core Programme and the Grid Hey, T., & Trefethen, A. E. International Conference on Computational Science (pp. 3-21). Berlin, Heidelberg: Springer Berlin Heidelberg (https://doi.org/10.1016/S0167-739X(02)00082-1)

117 Hey T. 2005. e-Science and open access. See https://www.researchgate.net/publication/28803295_E-Science_and_Open_Access (accessed 7 May 2024)

118 The Alan Turing Institute. Research. See https://www.turing.ac.uk/research (accessed 21 December 2023).

119 Arizona State University - School of Sustainability. See https://schoolofsustainability.asu.edu/ (accessed 21 December 2023).

120 The Royal Society. 2023 Science in the metaverse: policy implications of immersive technologies. See https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/ (accessed 21 December 2023).

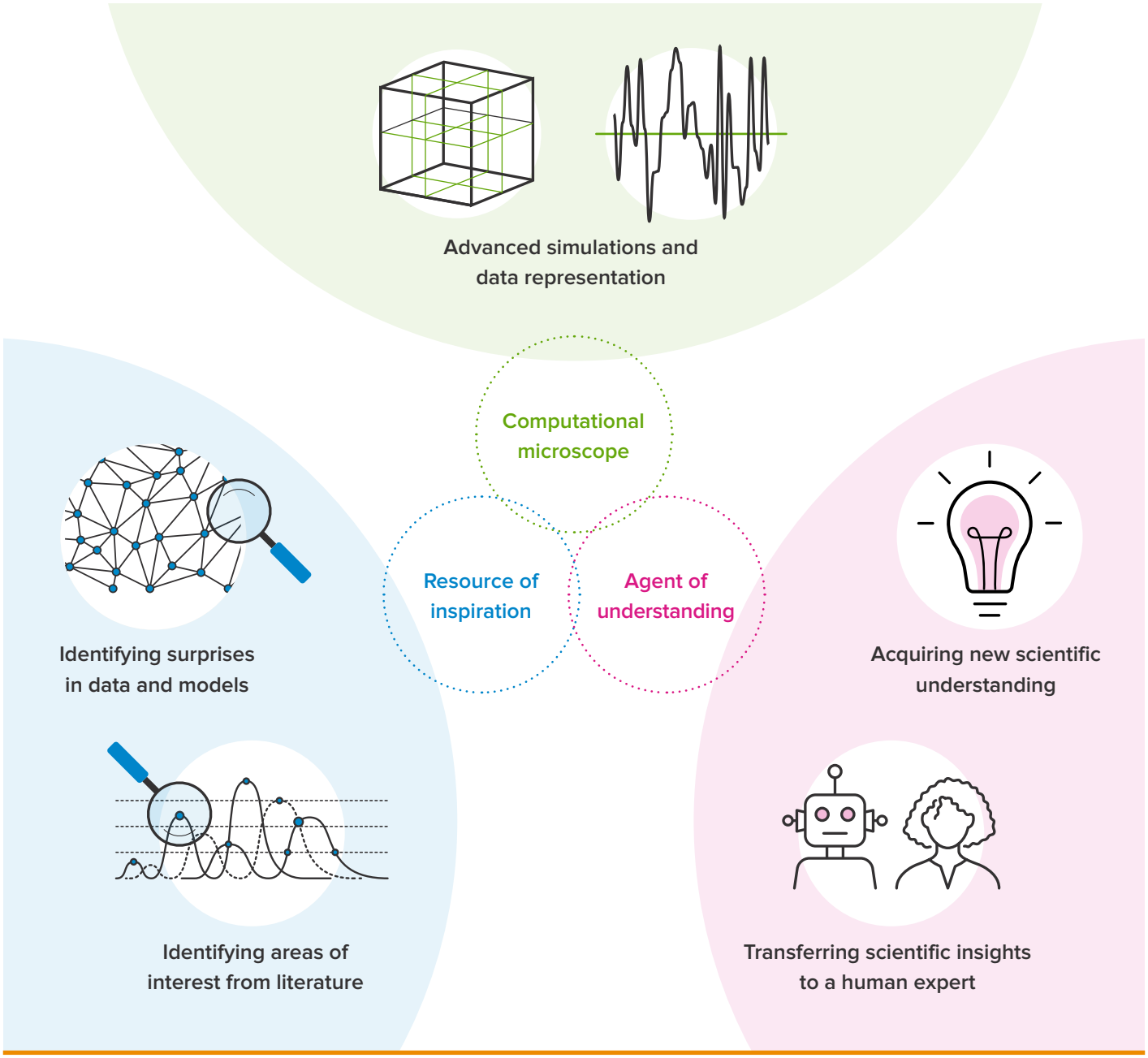
121 *Ibid*.

122 Krenn M. *et al.* 2022. On scientific understanding with artificial intelligence. *Nature Reviews Physics* **4**. (https://doi.org/10.1038/s42254-022-00518-3)

123 Messeri L, Crockett MJ. 2024 Artificial intelligence and illusions of understanding in scientific research. Nature. Mar;627(8002):49-58. (https://doi.org/10.1038/s41586-024-07146-0.)

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 understanding[124]



124 The diagram describes three possible ways in which AI can contribute to scientific understanding. The 'computational microscope' refers to the role of AI in providing information through advanced simulation and data representation that cannot be obtained through experimentation. 'Resource of inspiration' refers to scenarios in which AI provides information that expands the scope of human imagination or creativity. The 'agent of understanding' illustrates a scenario in which autonomous AI systems can share insights with human experts by translating observations into new knowledge. As of yet, there is no evidence to suggest that computers can act as true agents of scientific understanding. See: Krenn M. *et al*. 2022. On Scientific Understanding with Artificial Intelligence.

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

Pete Buttigieg
Participant in the US-UK Scientific Forum on Researcher Access to Data

AI and access to high-quality data

Data is the foundation of AI systems. The expression and principle of 'garbage in, garbage out', which dates to the early days of computing[125] and the writings of Charles Babbage FRS[126], remain applicable today. Poor-quality data which is incomplete, incorrect, or unrepresentative, can result in misleading outcomes. Ensuring high-quality datasets to train AI systems involves addressing challenges such as trust, access, bias, availability, and interoperability across the data lifecycle.

Drawing upon the 2023 US-UK Scientific Forum on Researcher Access to Data, organised by the Royal Society and the US National Academy of Sciences[127]; and interviews, the following themes emerged as key challenges associated with the use of data for AI-based scientific research:

1. Volume and heterogeneity

Data collected for scientific experiments can be extremely large, measuring into terabytes, petabytes, and exabytes in size[128]. This volume challenge applies to diverse fields including genomics, high-energy physics, climate science, and astronomy.

An example presented at the US-UK Scientific Forum[129] is the Event Horizon Telescope (Georgia Institute of Technology) which took the first two photographs of black holes.[130] The project, which involved ten telescopes across the world, recorded one petabyte of data per night. The vast size of this type of data and its often heterogenous nature makes objectives such as interoperability difficult to achieve. This challenge calls for integrated and central repositories that provide long-term stewardship and access for researchers, near real time dissemination and analysis, and solutions for the annotation of heterogeneous training data. For example, the National Oceanic and Atmospheric Administration are experimenting with a videogame solution called FathomVerse, asking players to identify broad categories of species they can see in video footage from exploration vessels[131] The establishment of large data infrastructures requires further consideration of its maintenance and environmental impact[132].

125 Mellin W. 1957 Work with new electronic 'brains' opens field for army math experts. *The Hammond Times*.

126 Babbage C. 1964. Passages from the life of a philosopher. Cambridge, UK: Cambridge University Press.

127 National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/27520.

128 Today, a typical PC or laptop comes with one terabyte storage. Petabytes are akin to the storage capacity of a thousand of these PCs and exabytes are akin to the storage capacity of a million.

129 National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/27520.

130 Event Horizon Telescope. See https://eventhorizontelescope.org/ (accessed 21 December 2023).

131 National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/27520.

132 SKAO. Portuguese prove SKA green energy system. See: https://www.skao.int/en/impact/440/portuguese-prove-ska-green-energy-system (accessed 21 March 2024)

2. The role of data institutions

Data institutions (eg archives, statistical agencies, data repositories) can play an important role in facilitating researcher access to data. These institutions have a range of functions, including:

- Protecting sensitive data and granting access under restricted conditions.
- Combining or linking data from multiple sources and providing insights and other services back to those who have contributed data.
- Creating open datasets that anyone can access, use, and share to further a particular mission or cause.
- Acting as a gatekeeper for data held by other organisations.
- Developing and maintaining identifiers, standards, and other infrastructure for a sector or field, such as by registering identifiers or publishing open standards.
- Enabling people to take a more active role in stewarding data about themselves and their communities.

Their fundamental role in maintaining the quality of data available for scientific research makes them akin to foundational infrastructure for AI tools. As such, they cost money and require maintenance.[133] Examples of data institutions include the US Government's open data website, Data.gov[134], the University of Michigan's Inter-university Consortium for Political and Social Research (holding data on more than 19,000 social and behavioural science studies)[135], and the UK's Office for National Statistics[136].

3. Sensitive data sharing

Limits to sharing sensitive data can block potential scientific breakthroughs. In the field of health care, for example, sharing and processing health data is complex due to its confidential and fragmented nature (both within institutions and across borders). While advancements in medical imaging, text, audio, and AI offer new possibilities for diagnosis and treatment, they also risk exposing sensitive patient information through imaging and metadata. Similarly, bad actors can also weaponise environmental data (eg rainfall, deforestation, or poaching data) to cause national security and environmental threats. Researchers are calling for a definition of 'sensitive data', that considers how data subject to exploitation, misuse, and misinterpretation[137] can cause societal and environmental harm[138].

133 National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/27520.

134 Data.gov. See https://data.gov/ (accessed 21 December 2023).

135 ICPSR. See https://www.icpsr.umich.edu/web/pages/index.html (accessed 21 December 2023).

136 Office for National Statistics. See https://www.ons.gov.uk/ (accessed 21 December 2023).

137 Leonelli S, Williamson H. 2022 Introduction: Towards Responsible Plant Data Linkage. In: Towards Responsible Plant Data Linkage: Data Challenges for Agricultural Research and Development. Springer International Publishing (https://doi.org/10.1007/978-3-031-13276-6_1).

138 National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/27520.

The use of trusted research environments and privacy enhancing technologies (including AI-based approaches such as federated machine learning), is enabling researchers to model problems without requiring data access, offering a potential technical solution to addressing concerns surrounding sensitive data. These are explained in detail in the Royal Society's 2019 report *Protecting privacy in practice*[139] and the 2023 report *From privacy to partnership* (which contains various use cases).[140]

Public trust and acceptability around the use of sensitive datasets relating to people (eg health information, demographics, location, etc.) is also essential. As set out in the Royal Society's 2023 report, *Creating resilient and trusted data systems*, trust in data sharing requires clarity of purpose and transparency in data flows, as well as robust systems for security and privacy[141]. Private sector actors such as IBM, Microsoft and Siemens are addressing publics concerns by establishing communities of trust[142]. Other approaches include data governance frameworks that encourage the public to get involved in data-driven scientific projects while retaining control of their data (eg data donation drives[143]).

139 The Royal Society. 2019 Protecting privacy in practice. See https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/ (accessed 21 December 2023).

140 *Ibid*.

141 The Royal Society. 2023 Creating resilient and trusted data systems. See https://royalsociety.org/topics-policy/projects/data-for-emergencies/ (accessed 21 December 2023).

142 Charter of Trust. See: www.charteroftrust.com (accessed 21 December 2023)

143 The Tidepool Big Data Donation Project. See: https://www.tidepool.org/bigdata (accessed 21 December 2023)

CASE STUDY 1

AI and rare disease diagnosis

A rare disease is a condition that affects fewer than 1 in 2,000 people and is often characterised by diverse, complex, and overlapping genetic manifestations[144]. Of the more than 7,000 rare diseases described worldwide, only 5% have a treatment[145]. A lack of understanding of underlying causes, fragmented patient data, and inadequate policies have contributed to making the diagnosis and treatment of rare diseases a public health challenge[146].

The application of ML and generative AI techniques offers an opportunity to overcome some of these limitations. Rare disease researchers are using ML techniques to analyse high-dimensional datasets, such as high-dimensional molecular data, to identify relevant biomarkers for known diseases or to identify new diseases[147]. The shift towards digitising health records is also creating opportunities to identify patients with rare diseases more promptly. Promising applications show potential to improve low diagnostic rates, treatments, and drug development processes[148].

AI applications in the field of rare diseases

- Leveraging medical imaging for early diagnosis: Clinicians are using AI to find patterns in large datasets of patient information, including genetic data and clinical records, that may indicate the presence of a rare disease. ML is particularly useful to analyse multimodal data from different sources, including imaging data (eg, MRI, X-rays) that is becoming standard practice to understand disease manifestation[149]. For example, researchers at the Institute for Genomics Statistics and Bioinformatic at the University of Bonn are using deep neural networks (DNNs) and computational facial analysis to accelerate the diagnosis of ultra-rare and novel disorders[150].
- Improving capabilities for automated diagnosis: ML techniques can also be used to improve automated diagnostic support for clinicians. Applying ML to very large multi-modal health datasets, such as UK Biobank[151], for example, is creating new possibilities to discover unknown and novel variants that can contribute to a molecular diagnosis of rare diseases[152].

144 Department of Health and Social Care. 2021. The UK Rare Diseases Framework. See: https://www.gov.uk/government/publications/uk-rare-diseases-framework/the-uk-rare-diseases-framework (accessed 30 September 2023).

145 Brasil S, Pascoal C, Francisco R, Dos Reis Ferreira V, Videira PA, Valadão AG. 2019 Artificial Intelligence (AI) in Rare Diseases: Is the Future Brighter? (https://doi.org/10:978 10.3390/genes10120978)

146 Decherchi S, Pedrini E, Mordenti M, Cavalli A, Sangiorgi L. 2021 Opportunities and challenges for machine learning in rare diseases. Frontiers in Medicine, 8, 747612. (https://doi.org/10.3389/fmed.2021.747612)

147 Banerjee J *et al*. 2023 Machine learning in rare disease. Nat Methods 20, 803–814. (https://doi.org/10.1038/s41592-023-01886-z)

148 Schaefer J, Lehne M, Schepers J, Prasser F, Thun S. 2020 The use of machine learning in rare diseases: a scoping review. Orphanet J Rare Dis.(https://doi.org/10.1186/s13023-020-01424-6)

149 *Ibid.*

150 Hsieh TC, Krawitz PM. 2023 Computational facial analysis for rare Mendelian disorders. American Journal of Medical Genetics Part C: Seminars in Medical Genetics. (https://doi.org/10.1002/ajmg.c.32061

151 UK Biobank. See: https://www.ukbiobank.ac.uk/ (accessed 21 December 2023)

152 Turro E *et al*. 2020 Whole-genome sequencing of patients with rare diseases in a national health system. Nature 583, 96–102 (https://doi.org/10.1038/s41586-020-2434-2

Similarly, ML is applied to determine whether data in health datasets can be used to identify patients who have not been previously tested for rare diseases and may have gone undiagnosed[153,154].

- Accelerate treatment and drug discovery: ML models, in particular generative AI, can be leveraged to accelerate the drug discovery process. Models screen molecular libraries to predict potential drug candidates and assess their effectiveness in treating specific rare diseases.[155] This area tends to be dominated by private sector pharmaceutical companies such as Insilico Medicine, Recursion, or Healx[156].

Data challenges for the application of AI for rare disease studies

- Limited data availability: Rare diseases affect a very small percentage of the global population. Relevant data – if available – is siloed, scattered, behind paywalls or commercially owned. This scarcity of data can make it difficult to train accurate and robust AI models[157]. This is exacerbated by a lack of channels to coordinate across labs and institutions to integrate and cross reference datasets.
- Biased and unrepresentative datasets: When rare disease data is available, it can be unrepresentative, creating issues which span from potential false positives or negatives to the underrepresentation of different age groups or ethnic minorities. Imbalanced data can lead to biased and 'overfitted' models that rely on patterns that are unique to the training data and thereby perform poorly on new datasets, with implications for transferability and generalisability of ML models across contexts.
- Heterogenous and noisy data: The use of ML is best suited for large and well-curated datasets, while rare disease data can be heterogenous, incomplete, or incorrectly labelled (eg, misdiagnoses or incorrect labelling can be common in rare disease studies due to a limited or evolving understanding of a condition)[158]. Data related to rare diseases may come from various sources, including clinical records, genetic testing, and patient surveys. These data sources may have different formats, quality standards, and levels of detail. Integrating and harmonising such heterogeneous data can be a significant challenge.

153 Cohen A M *et al*. Detecting rare diseases in electronic health records using machine learning and knowledge engineering: Case study of acute hepatic porphyria. *PLoS.* (https://doi.org/10.1371/journal.pone.0238277)

154 Hersh W R *et al*. 2022 Clinical study applying machine learning to detect a rare disease: results and lessons learned. *JAMIA Open,* **5**. (https://doi.org/10.1093/jamiaopen/ooac053)

155 Nag S, *et al*. 2022 Deep learning tools for advancing drug discovery and development. *3 Biotech*. **12:** 110.

156 Steve Nouri. Generative AI Drugs Are Coming. See: https://www.forbes.com/sites/forbestechcouncil/2023/09/05/generative-ai-drugs-are-coming/ (accessed September 30 2023)

157 Banerjee J *et al*. 2023 Machine learning in rare disease. Nat Methods 20, 803–814. (https://doi.org/10.1038/s41592-023-01886-z)

158 *Ibid.*

- Cost and resource constraints: Collecting and annotating data for rare diseases can be expensive and time-consuming. Many healthcare organisations, including in resource-limited settings, may not have access to resources to invest, build and maintain large-scale datasets for rare diseases[159].
- Sensitive data: Medical data is highly sensitive, and there are strict data governance, management, and protection considerations. Anonymising and de-identifying data is a common practice, however small samples in rare diseases increase the likelihood of identifying people through triangulation[160]. Sharing rare disease data while ensuring patient privacy and complying with regulations can be complex. More resources are needed to set up trusted and secure research environments that enable sensitive data sharing.

Strategies to maximise the value of AI in rare disease research

Concerted efforts in data sharing, standardisation and data governance can pave the way for AI to make a significant impact on the study of rare diseases and improving the lives of those affected by these conditions.

- Independent and interoperable patient registries: Cross-institutional, international collaboration is needed to create large, centralised datasets suitable for ML-based research[161]. Data pooling initiatives are also desirable, such as regional or global patient registries[162] that can widen access to relevant data and promote standardisation and interoperability of registries across institutions. For example, the European Rare Diseases Platform has released the 'Set of common data elements for Rare Diseases Registration'. Establishing federated learning infrastructures, such as Gaia-X, can also facilitate sensitive data sharing[163].
- Use AI for data augmentation: Generative techniques can be an effective way to address data scarcity, noise, or incompleteness[164]. For example, data augmentation strategies or the use of synthetic data can be used to populate incomplete datasets with artificial samples that increase diversity and representativeness of datasets (eg, address outliers and biases), minimising the need for personal data. Similarly, computer vision approaches can be used to improve the quality and fidelity of imaging data. Outstanding challenges include ensuring the reliability and adequate training of generative models.
- Establish multi-stakeholder cooperation networks: Rare disease researchers stressed the importance of collaboration to widen access to resources and align multiple stakeholder interests. The Asia-Pacific Economic Cooperation's Rare Disease Network is a relevant model that brings together policymakers, academia, and industry to manage and harmonise data practices[165].

159 The Royal Society interviews with scientists and researchers. 2022 – 2023

160 The Royal Society interviews with scientists and researchers. 2022 – 2023

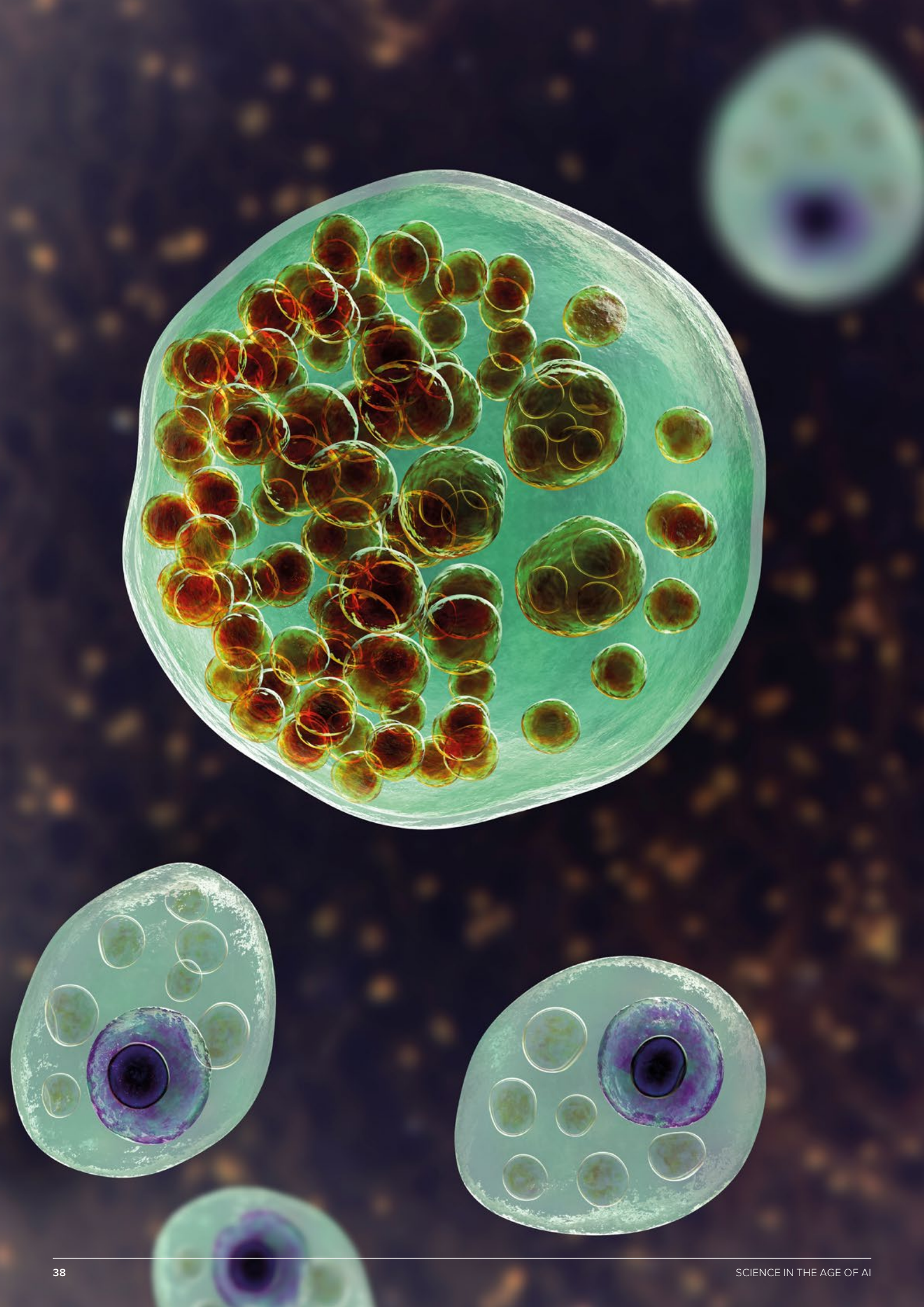
161 Boycott KM *et al*. 2017 International cooperation to enable the diagnosis of all rare genetic diseases. Am. J. Hum. Genet. 100, 695–705. (https://doi.org/10.1016/j.ajhg.2017.04.003)

162 Bellgard MI, Snelling T, McGree JM. 2019 RD-RAP: beyond rare disease patient registries, devising a comprehensive data and analytic framework. Orphanet J Rare Dis 14, 176. (https://doi.org/10.1186/s13023-019-1139-9)

163 Decherchi S, Pedrini E, Mordenti M, Cavalli A, Sangiorgi,L. 2021 Opportunities and challenges for machine learning in rare diseases. Frontiers in Medicine, 8, 747612 (https://doi.org/10.3389/fmed.2021.747612)

164 Kokosi T, Harron K. 2022. Synthetic data in medical research. BMJ medicine, 1. (https://doi.org/10.1136/bmjmed-2022-000167)

165 Global Rare Disease Policy Network. See: https://www.rarediseasepolicy.org/ (accessed 21 March 2024)



Chapter two
Research integrity and trustworthiness

Left
Rhinosporidium seeberi parasite, the causative agent of rhinosporidiosis.

Research integrity and trustworthiness

"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 crisis, thus averting a complete loss of trust in scientific expertise by civil society."

Professor Sabina Leonelli[166]

Trust in AI is essential for its responsible use in scientific research, particularly as scientists become increasingly reliant on these technologies[167]. This reliance hinges on an assumption that AI-based systems – as well as their analysis and outputs – can produce reliable, low-error, and trustworthy findings.

However, the adoption of AI in scientific research has been coupled with challenges to rigour and scientific integrity. Core issues include a lack of understanding about how AI models work, insufficient documentation of experiments, and scientists lacking the required technical expertise for building, testing and finding errors in a model. A growing body of irreproducible studies using ML techniques are also raising concerns regarding the challenges to reproduce AI-based experiments and the reliability of AI-based results and discoveries[168]. Together, these issues pose risks not just to science, but also to society if the deployment of unreliable or untrustworthy AI technologies leads to harmful outcomes[169].

Based on interviews and a roundtable on reproducibility conducted for this report, the following observations capture unique challenges AI poses for research integrity and trustworthiness.

Reproducibility challenges in AI-based research

Reproducibility refers to the ability of independent researchers to scrutinise the results of a research study, replicate them, and reproduce an experiment in future studies[170].

If researchers develop an overreliance on AI for data analysis, while remaining unable to explain how conclusions were reached and how to reproduce a study[171], it will not meet thresholds for scrutiny and verification. Similarly, if results cannot be verified, they can contribute to inflated expectations, exaggerated claims of accuracy, or research outputs based on spurious correlations[172]. In the case of AI-based research, being able to reproduce a study not only involves replicating the method, but also being able to reproduce the code, data, and environmental conditions under which the experiment was conducted (eg computing, hardware, software)[173,174].

166 Leonelli S. 2018 Rethinking reproducibility as a criterion for research quality. In a symposium on Mary Morgan: "Curiosity, Imagination, and Surprise" of Research in the History of Economic Thought and Methodology. *Emerald Publishing Limited*. January 2018. **36B**, 129-146 (**https://doi.org/10.1108/S0743-415420180000036B009**)

167 Echterhölter A, Schröter J, Sudmann A. 2021 How is Artificial Intelligence Changing Science? Research in the Era of Learning Algorithms. OSF Preprint (https://doi.org/10.33767/osf.io/28pnx)

168 Sohn E. 2023 The reproducibility issues that haunt health-care AI. Nature. 9 January 2023. See https://www.nature.com/articles/d41586-023-00023-2 (accessed 21 December 2023)

169 Sambasivan N *et al*. 2021 "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. (https://doi.org/10.1145/3411764.3445518)

170 Haibe-Kains B *et al*. 2020 Transparency and reproducibility in artificial intelligence. Nature. 586, E14–E16. (https://doi.org/10.1038/s41586-020-2766-y)

171 Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See https://royalsociety.org/current-topics/ai-data/ (accessed 7 May 2024)

172 Echterhölter A, Schröter J, Sudmann A. 2021 How is Artificial Intelligence Changing Science? Research in the Era of Learning Algorithms. OSF Preprint (https://doi.org/10.33767/osf.io/28pnx)

173 Gundersen O, Gil Y and Aha D. 2018 On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI Magazine*. **39**: 56-68. (doi.org/10.1609/aimag.v39i3.2816)

174 Gunderson O, Coakley K, Kirkpatrick C, and Gil Y. 2022 Sources of irreproducibility in machine learning: A review. *arXiv preprint*. (doi.org/10.48550/arXiv.2204.07610)

Reproducibility failures not only risk the validity of the individual study[175], but can also affect research conducted for other studies, including those in other disciplines. For example, a study led by the Center for Statistics and Machine Learning at Princeton University showed how 'data leakage' in one study (a leading cause of errors in ML applications due to errors in training data or model features) may affect 294 papers across 17 scientific fields, including high-stakes fields like medicine[176]. Furthermore, these types of issues are likely to be underreported due to factors such as unpublished data; insufficient documentation; absence of mechanisms to report failed experiments; and high variability across experimentation or research contexts[177].

Opacity and the black-box nature of machine learning

At the core of the reproducibility challenge are opaque ML-based models that not every scientist can explain, interpret, or understand. ML models are commonly referred to as 'black-box models' (models that can produce useful information and outputs, even when researchers do not understand exactly how the system works). The opaque nature of models limits explainability and the ability of scientists to interpret how ML models arrive at specific results or conclusions[178].

Explainable AI (See Box 1) can help researchers identify errors in data, models, or assumptions – mitigating challenges such as data bias – and ensure these systems produce high quality results which can be used for real-world implementation[179]. This can become a significant challenge for scientists who integrate highly variable and complex models into their work, such as deep learning models, that are known to outperform less complex and more linear and transparent models.

Opacity increases when models are developed in a commercial setting. For instance, most leading LLMs are developed by large technology companies like Google, Microsoft, Meta, and OpenAI. These models are proprietary systems, and as such, reveal limited information about their model architecture, training data, and the decision-making processes that would enhance understanding[180].

"There may be a disproportionate problem with machine learning. We've come very far with the ability to handle huge amounts of data, using software that is very competent and well developed. But I think perhaps a lot of people using it don't actually understand what they're doing in a way that may not be so true for other areas. It's a compounded problem, where there are many, many things you can get wrong. I wonder how many people really understand the software that they're using."

Royal Society roundtable participant

175 McDermott M, Wang S, Marinsek N, Ranganath R, Foschini L, and Ghassemi M. 2021 Reproducibility in machine learning for health research: Still a way to go. *Sci. Transl. Med.* **13**, eabb1655. (doi.org/10.1126/scitranslmed.abb1655)

176 Kapoor S and Narayanan A. 2023 Leakage and the reproducibility crisis in machine-learning-based science. *Patterns.* 4(9) (doi.org/10.1016/j.patter.2023.100804)

177 Gundersen O, Gil Y and Aha D. 2018 On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI Magazine*. **39**: 56-68. (doi.org/10.1609/aimag.v39i3.2816)

178 Royal Society. Royal Society response on Reproducibility and Research Integrity. See: https://royalsociety.org/news-resources/publications/2021/research-reproducibility/ (accessed 7 March 2024)

179 The Royal Society. 2019 Explainable AI: the basics. See https://royalsociety.org/topics-policy/projects/explainable-ai/ (accessed 21 December 2023).

180 Bommasani *et al.* 2021. On the opportunities and risks of foundation models. See: https://crfm.stanford.edu/assets/report.pdf (accessed March 21 2024)

BOX 1

Explainability and interpretability

Explainability and interpretability refer to information that allows users to understand how an AI system works and the reasoning behind its outputs[181]. 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?"[182].

As set out in the Royal Society's 2019 report, *Explainable AI: The basics*[183], 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 discoveries[184].
- 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 models[185].
- 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 models[186].

181 Marcinkevičs, R., Vogt, J. E. 2023. Interpretable and explainable machine learning: A methods-centric overview with concrete examples (https://doi.org/10.1002/widm.1493)

182 Lipton,. 2018. The mythos of model interpretability. Queue, 16(3), 31–57. (https://doi.org/10.1145/3236386.3241340)

183 The Royal Society. 2019 Explainable AI: the basics. See https://royalsociety.org/topics-policy/projects/explainable-ai/ (accessed 21 December 2023).

184 Li Z, Ji J, and Zhang Y. 2023 From Kepler to Newton: Explainable AI for science. *arXiv preprint*. (doi.org/10.48550/arXiv.2111.12210)

185 McDermid J, Jia Y, Porter Z and Habli I. 2021 Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A*. 379(2207), 20200363. (doi.org/10.1098/rsta.2020.0363)

186 McGough M. 2018 How bad is Sacramento's air, exactly? Google results appear at odds with reality, some say. *Sacramento Bee*. 7 August 2018. See https://www.sacbee.com/news/california/fires/article216227775.html (accessed 21 December 2023).

The trade-off between explainability and performance

One of the main limitations to explainability is what has been referred to as a trade-off between explainability and accuracy[187]. It is considered that the highest accuracy of modelling (in terms of prediction or classification) for large modern datasets is often achieved by opaque complex models. This is coupled with a general acceptance among AI users of opacity involved in using ML models. The competitive and fast-paced adoption of AI in scientific research is entrenching this acceptance even further. The current AI ecosystem rewards high performance and competitive models that are 'useful' and accurate, rather than transparent, accessible, or 'user-friendly'[188].

These perceptions raise questions on whether opacity has been normalised as the new status quo and whether explainability is a feasible goal worth pursuing[189,190]. Furthermore, while transparency can enhance understanding, providing complex technical information about a system may not always improve the ability of end users to interact with and understand systems[191]. As an alternative to restricting research to explainable models only, some have suggested a greater focus on improving the interpretability of models[192].

Promising approaches to improve explainability or interpretability include:

- Discipline-specific explainable AI methods (XAI): As users from diverse disciplines integrate AI into their work, explainability methods are becoming discipline and application-specific. XAI methods are emerging in fields such as material science[193], biomedicine[194], earth science[195],

187 Lundberg S and Lee S. 2017 A unified approach to interpreting model predictions. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*. 4768–4777. (dl.acm.org/doi/10.5555/3295222.3295230)

188 Cartwright H. 2023 Interpretability: Should – and can – we understand the reasoning of machine-learning systems? In: OECD (ed.) *Artificial Intelligence in Science*. OECD. (doi.org/10.1787/a8d820bd-en)

189 Royal Society roundtable on reproducibility, April 2023.

190 Birhane A *et al.* 2023 Science in the age of large language models. *Nat Rev Phys* **5**, 277–280. (doi.org/10.1038/s42254-023-00581-4)

191 Bell A, Solano-Kamaiko I, Nov O, and Stoyanovich J. 2022 It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. *In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. 248–266. (doi.org/10.1145/3531146.3533090)

192 Miller K. 2021 Should AI models be explainable? That depends. *Stanford University Human-Centered Artificial Intelligence*. See: https://hai.stanford.edu/news/should-ai-models-be-explainable-depends (accessed 21 December 2023).

193 Zhong X *et al.* 2022 Explainable machine learning in materials science. *NPJ Comput Mater* 8, 204. (doi.org/10.1038/s41524-022-00884-7)

194 Combi C *et al.* 2022 A manifesto on explainability for artificial intelligence in medicine. *Artificial intelligence in medicine*. 133, 102423. (doi.org/10.1016/j.artmed.2022.102423)

195 Hanson, B. Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research. See https://www.nature.com/articles/d41586-023-03316-8 (accessed 5 March 2024)

"One of the things that is true with modelling is you can get almost any result you want based on the assumptions you use to drive them. I think this is a dangerous area that our field is moving in. It's too much reliance on model results and the pretty pictures that come out of it as a reproduction of truth."

Royal Society roundtable participant

and environmental research[196] for multiple purposes. These include enhancing scientific understanding derived from AI (eg better understanding of physical principles and generation of new hypothesis[197]); improving oversight and enforcement of environmental protection regulations; and minimising the environmental footprint of AI systems[198].

- Glass-box architectures: Glass-box model architectures aim to make LLM's internal data representations more transparent by incorporating attention mechanisms, modular structures, and visualisation tools that can help surface how information flows through layers of the neural network. In addition, augmented training techniques like adversarial learning and contrastive examples can probe the model's decision boundaries. Analysing when the LLM succeeds or fails on these special training samples provides insights into its reasoning process[199,200].
- Knowledge graphs: Knowledge graphs are an advanced data structure that represent information in a network of interlinked entities. They reduce reliance on opaque statistical patterns in training data for LLMs. Medical LLMs, for example, can leverage ontological biomedical data in knowledge graphs for transparent structured reasoning about diseases and treatments. During inference, LLMs consult knowledge graphs for relevant facts, providing a grounded framework alongside their intrinsic pattern recognition. Joint training with knowledge graphs improves LLMs' factual reasoning and aids in identifying gaps or misconceptions through audits[201].

Barriers limiting reproducibility

Beyond technical challenges, there are a series of institutional and social constraints that prevent researchers from adopting more rigorous and transparent processes. Table 1 lists key barriers to reproducibility in AI-based research[202].

196 Arashpour M. 2023 AI explainability framework for environmental management research. *Journal of environmental management*. 342, 118149. (doi.org/10.1016/j.jenvman.2023.118149)

197 Zhong X *et al.* 2022 Explainable machine learning in materials science. *npj Comput Mater* **8,** 204. (doi.org/10.1038/s41524-022-00884-7)

198 Arashpour M. 2023 AI explainability framework for environmental management research. *Journal of environmental management*. 342, 118149. (doi.org/10.1016/j.jenvman.2023.118149)

199 Lengerich, B J *et al.* 2023. LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs. arXiv preprint (https://doi.org/10.48550/arXiv.2308.01157)

200 Garrett BL, Rudin C 2023. The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice. Cornell Law Review, Forthcoming, Duke Law School Public Law & Legal Theory Series.

201 Gaur, M, Faldu, K, & Sheth, A 2021. Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Computing*, **25**, 51-59.

202 Royal Society roundtable on reproducibility, April 2023.

TABLE 1

Barriers to reproducibility and examples

Barrier to reproducibility	Examples
Misconceptions and assumptions about ML[203]	• 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	• 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 models[204]. • 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 cannot[205].
Documentation and transparency practices	• 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	• 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	• 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 possible[206].

203 Benjamin D *et al.* 2018 Redefine statistical significance. *Nat Hum Behav* 2, 6–10. (doi.org/10.1038/s41562-017-0189-z)

204 Bommasani *et al.* 2021. On the opportunities and risks of foundation models. See: https://crfm.stanford.edu/assets/report.pdf (accessed March 21 2024)

205 *Ibid.*

206 Rudin C. 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1, 206–215). (doi.org/10.1038/s42256-019-0048-x)

"The [reproducibility] crisis or the challenge with machine learning is about who is involved in determining how reproducibility is defined and for whom will this definition be useful? Are we applying contextual knowledge and situated understanding of what the technology will be used for? To what extent will it support diverse research communities to enhance the trustworthiness of their experiments?"

Royal Society roundtable participant

Research culture and reproducibility

The 'publish or perish' culture was highlighted as a key limiting factor for scientists, in so far as it rewards the number and type of publications as requisite for career progression but does not recognise datasets, documentation or reproduced studies as outputs worth rewarding. Coupled with ineffective practices of quality control and self-correction within journals, the current publishing system is considered to play a significant role in diminishing incentives to conduct the time-intensive and collaborative work required to demonstrate reproducibility[207].

These challenges are further explored in the Royal Society's 2018 report, *Research culture: Embedding inclusive excellence*. The report highlights that one of the primary incentives for disseminating research findings should be to benefit the community as a whole and to advance the research enterprise[208]. It also discusses the value of transparency for embedding a culture of integrity and as a means for guarding against unnecessary duplication of research.

Reproducibility across contexts

A universal approach to achieving reproducibility is not desirable. This usually demands a high level of control over environmental and social conditions of a study, as well as a direct replication of inputs, outputs, and methods that may be unfeasible across diverse research environments, cultures, and contexts.

A standardised approach to reproducibility can also discourage researchers from approaching documentation and reporting from a reflexive standpoint that addresses variability and the more idiosyncratic aspects of scientific research[209]. Models that are not generalisable across contexts can, for example, offer valuable insights regarding the source of variations and why they matter. In the context of healthcare, the local conditions (eg, admission protocols, lab testing, record managements, or clinician-patient interactions) of a hospital can significantly shift the outputs.

An alternative to a standardised approach to improving reproducibility is a more contextual approach to documentation and research protocols that embraces variability and provides insight into the local adaptation of models across contexts[210]. This approach has the potential to support researchers who wish to adapt models, rather than exporting or importing models that do not transfer well to different geographical and cultural 'contexts of discovery'[211].

207 Leonelli S. 2018 Rethinking reproducibility as a criterion for research quality. In a symposium on Mary Morgan: "Curiosity, Imagination, and Surprise" of Research in the History of Economic Thought and Methodology. *Emerald Publishing Limited*. January 2018. **36B**, 129-146 (https://doi.org/10.1108/S0743-41542018000036B009)

208 The Royal Society. 2018 Research culture: embedding inclusive excellence. See https://royalsociety.org/topicspolicy/publications/2018/research-culture-embedding-inclusive-excellence/ (accessed 21 December 2023)

209 Leonelli S. 2018 Rethinking reproducibility as a criterion for research quality. In a symposium on Mary Morgan: "Curiosity, Imagination, and Surprise" of Research in the History of Economic Thought and Methodology. *Emerald Publishing Limited*. January 2018. **36B**, 129-146 (https://doi.org/10.1108/S0743-41542018000036B009)

210 Miller K. 2022 Healthcare algorithms don't always need to be generalizable. *Stanford University Human-Centered Artificial Intelligence*. See https://hai.stanford.edu/news/healthcare-algorithms-dont-always-need-be-generalizable (accessed 21 December 2023).

211 Leonelli S. 2018 Rethinking reproducibility as a criterion for research quality. In a symposium on Mary Morgan: "Curiosity, Imagination, and Surprise" of Research in the History of Economic Thought and Methodology. *Emerald Publishing Limited*. January 2018. **36B**, 129-146 (https://doi.org/10.1108/S0743-41542018000036B009)

BOX 2

Robustness and generalisability in machine learning

Model robustness refers to a model's ability to perform accurately across contexts[212]. While modern ML models are optimised to accomplish narrow and specific tasks, developments in AI involve developing the capacity for 'generalising' or transferring learning from a training task to novel applications. Without generalisability, AI models might perform well on the training data but fail to perform well on new, unseen data, or in new research environments. Reproducibility plays a major role in ensuring parties from diverse research environments can replicate an experiment and test generalisability.

Advancing transparency and trustworthiness

Challenges with trustworthiness have led scientists to develop research protocols, standards, tools, and open science practices to ensure transparency and scientific rigour in AI-based research[213]. These include:

Incentivising the publication of reproducibility reports

- Pre-registration and registered reports. Initiatives to publish methodologies as promoted by the Centre for Open Science can enhance transparency of research studies, by encouraging researchers to document their research plan before conducting further research and submitting the methodology to peer review[214].
- Pre-print servers. The increased use of preprint servers (eg such as bioRxiv by the biological and biomedical communities) may play a role in facilitating communication of successful and unsuccessful replication results.
- Grand challenges. For example, the ML Reproducibility Challenge invites participants to reproduce papers published in eleven top ML conferences and publish a community-led reproducibility report documenting findings[215].

212 The Royal Society. 2017 Machine Learning: The power and promise of computers that learn by example. See https://royalsociety.org/topics-policy/projects/machine-learning/ (accessed 21 December 2023).

213 Birhane A *et al.* 2023 Science in the age of large language models. *Nat Rev Phys* **5**, 277–280. (doi.org/10.1038/s42254-023-00581-4)

214 Center for Open Science. What is preregistration? See https://www.cos.io/initiatives/prereg (accessed 21 December 2023).

215 Papers With Code. ML Reproducibility Challenge 2022. See https://paperswithcode.com/rc2022 (accessed 21 December 2023).

Guidance to produce documentation and follow open science practices

- Reproducibility checklists and protocols. Examples include the Machine Learning Reproducibility Checklist[216], Checklist for AI in Medical Imaging (CLAIM)[217], or the field-agnostic REFORMS checklist[218], developed by experts in computer science, mathematics, social science, and health research. These facilitate compliance and documentation of the multiple dimensions of reproducibility.
- Community standards for documentation. The development of domain-specific community standards such as TRIPOD-AI[219] provide guidance on how to document, report and reproduce machine-learning based prediction model studies in health research. The synthetic biology and genomics community have also defined experimental protocol standards and documentation of the genomic workflow to improve reproducibility[220,221].
- The release of data sheets and model cards. Industry can play an important role in releasing information that provide insight into what a model does; its intended audience; intended uses; potential limitations; confidence metrics; and information about the model architecture and the training data. Meta[222], Google[223], and Hugging Face[224] have released different iterations of model cards.
- Context-aware documentation. Involving diverse actors in defining how reproducibility is defined; promoting reporting mechanisms that explicitly address contextual inputs and sources of variation; and documenting how local or team culture influences implementation[225].

216 McGill School of Computer Science. The Machine Learning Reproducibility Checklist v2.0. See: https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf (accessed 21 December 2023).

217 Mongan J, Moy L, and Kahn C. 2020 Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiology. Artificial intelligence*, 2(2), e200029. (doi.org/10.1148/ryai.2020200029)

218 Reporting standards for ML-based science. See: https://reforms.cs.princeton.edu/ (accessed 21 December 2023).

219 Collins G *et al.* 2021 Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ open*, 11(7), e048008. (doi.org/10.1136/bmjopen-2020-048008)

220 Lin X. 2020 Learning Lessons on Reproducibility and Replicability in Large Scale Genome-Wide Association Studies. *Harvard Data Science Review*. **2**. (doi.org/10.1162/99608f92.33703976)

221 Kanwal S *et al.* 2017 Investigating reproducibility and tracking provenance – A genomic workflow case study. *BMC Bioinformatics* **18,** 337. (doi.org/10.1186/s12859-017-1747-0)

222 Meta. 2022 System Cards, a new resource for understanding how AI systems work. *Meta.* 23 February 2022. See https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/ (accessed 21 December 2023).

223 Google. Model Cards. See https://modelcards.withgoogle.com/about (accessed 21 December 2023).

224 Hugging Face. Model Cards. See https://huggingface.co/docs/hub/model-cards (accessed 21 December 2023).

225 Leonelli S. 2018 Rethinking reproducibility as a criterion for research quality. In a symposium on Mary Morgan: "Curiosity, Imagination, and Surprise" of Research in the History of Economic Thought and Methodology. *Emerald Publishing Limited*. January 2018. **36B**, 129-146 (**https://doi.org/10.1108/S0743-415420180000036B009**)

Collaborative and accessible tools and platforms

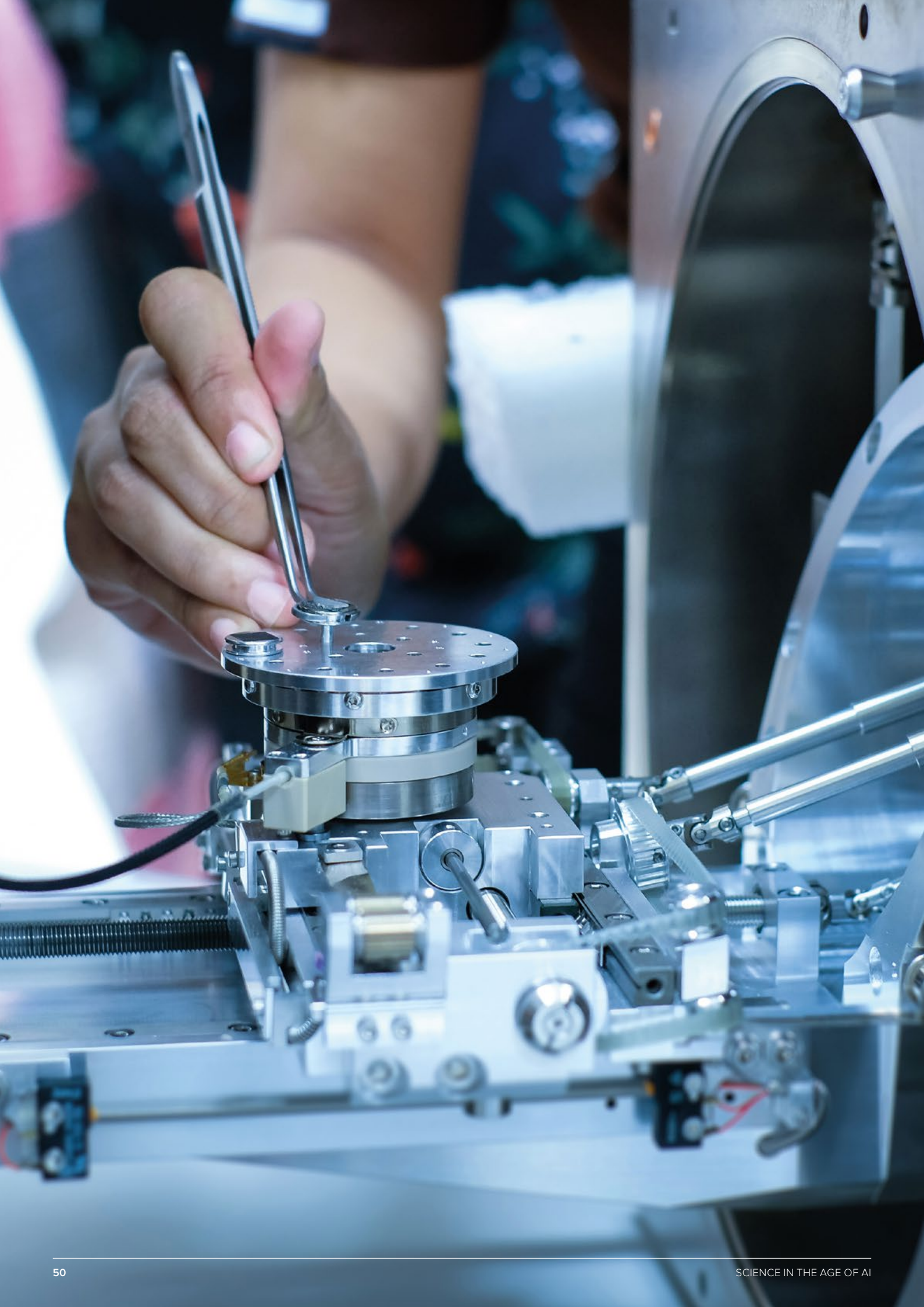
- Online collaborative platforms and repositories that facilitate the sharing of datasets; software versions; algorithms; workflows; and methods. Examples include CodaLab[226] and OpenML[227].
- Interactive and code-free tools. The development of accessible AI tools can help build trust and incorporate user expertise into the design and evaluation of model parameters. Examples include interactive dashboards[228] and code-free solutions that employ user-friendly interfaces to display datasets, confidence metrics and train the model with new examples. Code-free tools and 'edge models' can also be used in regions without internet access with limitations in functionality[229].

226 CodaLab. CodaLab Worksheets. See https://worksheets.codalab.org/ (accessed 21 December 2023).

227 OpenML. See https://www.openml.org/ (accessed 21 December 2023).

228 Morris M. 2023 Scientists' perspectives on the potential for generative AI in their fields. *Google Research*. (doi.org/10.48550/arXiv.2304.01420)

229 Korot E *et al.* 2021 Code-free deep learning for multi-modality medical image classification. *Nat Mach Intell*. **3**, 288–298. (doi.org/10.1038/s42256-021-00305-2)



Chapter three
Research skills and interdisciplinarity

Left
Preparation of nanomaterials for Scanning Electron Microscope (SEM) machine.

Research skills and interdisciplinarity

"We're getting to a scale of data and complexity, that you can't do it all yourself, or you can but you will be limited. If you want to be successful, you need to understand how to collaborate and co-create with people. And that is hard. It is painful and it is slower, but potentially more impactful, right? [...] If you want quick results, do it yourself. If you want impactful things, you need to work together and do things differently."

Royal Society roundtable participant

The successful application of AI in scientific research, and its translation to real-world value, may require interdisciplinary skills and knowledge[230]. Interdisciplinary research (IDR) involves activities that integrate more than one discipline with the aim to create new knowledge or solve a common problem[231]. Computer scientists need domain expertise to design suitable models, while domain experts need AI expertise to leverage those tools for their research. This interdisciplinary collaboration can facilitate knowledge sharing and drive innovative solutions to complex global problems such as climate change, biodiversity loss and epidemics[232,233].

This chapter draws upon insights obtained from a roundtable hosted by the Royal Society on the role of interdisciplinarity in ensuring the advancement of AI-based scientific research[234].

Challenges for interdisciplinarity in AI-based research

Interdisciplinary collaboration faces various barriers including the following:

1. **Siloed academic disciplines and research cultures**
   Disciplines often operate within distinct 'epistemic cultures' encompassing norms; methodologies; theoretical frameworks; evaluation; and funding models[235]. Establishing a shared language demands persistent effort and initiative to bridge terminological, paradigmatic, and cognitive gaps[236]. Interdisciplinary centres can take an interdisciplinary approach to AI-based research. For instance, the Centre for the Study of Existential Risk at the University of Cambridge studies the biological, environmental, global justice, and extreme technological risks of AI[237,238]. Similarly, doctoral training programmes, such as the UKRI Centre for Doctoral Training in Environmental Intelligence[239], can foster an interdisciplinary culture by providing PhD students with cross-disciplinary training in AI ethics, governance, and responsible innovation.

230 Eguíluz, V M, Mirasso, C R, & Vicente, R 2021. Fundamentals and Applications of AI: An Interdisciplinary Perspective. *Frontiers in Physics,* **8**. (https://doi.org/10.3389/fphy.2020.633494)

231 Weber, C T, & Syed, S. 2019. Interdisciplinary optimism? Sentiment analysis of Twitter data. *Royal Society open science,* **6**. (https://doi.org/10.1098/rsos.190473)

232 The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023.

233 Wang H *et al* 2023. Scientific discovery in the age of artificial intelligence. *Nature*, **620**. 47-60. (https://doi.org/10.1038/s41586-023-06221-2)

234 The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023.

235 Zeller F, Dwyer L. 2022 Systems of collaboration: challenges and solutions for interdisciplinary research in AI and social robotics. Discover Artificial Intelligence, 2.12. (https://doi.org/10.1007/s44163-022-00027-3)

236 The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, 2023.

237 University of Cambridge. Centre for the study of existential risk. See https://www.cser.ac.uk/ (accessed 13 December 2023)

238 The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023.

239 University of Exeter. Environmental Intelligence: Data Science & AI for Sustainable Futures. See https://www.exeter.ac.uk/research/eicdt/ (accessed 22 February 2024)

2. Siloed data infrastructures

While disciplines are collecting large amounts of data, siloed data infrastructure limits knowledge and data sharing[240]. Integrated data systems and consortiums can mitigate data redundancies, standardise data processing and facilitate researcher access to data, while reducing duplication of effort. Examples include the Environmental Data Service, a network providing UK environmental science data and tools for interdisciplinary analysis[241] and the Ocean Data Platform, which aggregates global ocean data in a cloud environment, overcoming storage and computing limitations[242].

3. Different publication models across scientific disciplines

Divergent publication models limit interdisciplinary collaboration in AI research Examples include disparities in publication venues (conference vs. journal publications), publication styles (single vs. group author publications), and levels of disclosure (open vs. closed science), impacting incentives and readiness to collaborate in AI-based research[243].

The emerging AI conference publication model, driven by frequent deadlines, is accelerating paper output leading to a cycle of frequent conferences, such as the Conference on Neural Information Processing Systems (NeurIPS) and the International Conference on ML (ICML)[244]. This may result in a high volume of papers that lack depth and quality, contrasting the iterative journal review process[245]. Secretive practices (eg closed data and models) also conflict with open science principles. This model differs significantly from traditional academic journal publications and may limit trustworthy IDR in AI-based research.

A hybrid journal-conference model where papers undergo thorough review in short turnaround journals, such as the Journal of Machine Learning Research (JMLR), before conference presentations could encourage higher-quality results and facilitate collaborations across disciplines valuing different publication norms[246]. Achieving this is likely to require a multi-stakeholder approach including funding agencies, research institutions, and AI conference communities to balance cutting-edge dissemination with deeper review.

240 National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/27520.

241 UKRI NERC Environmental Data Service. See: https://eds.ukri.org/

242 National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/27520.

243 The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023.

244 Bengio Y. 2020 Time to rethink the publication process in machine learning. See https://yoshuabengio.org/2020/02/26/time-to-rethink-the-publication-process-in-machine-learning/ (accessed 10 January 2024)

245 *Ibid.*

246 Slow Science. See http://slow-science.org/ (accessed 10 January 2024)

"Students are expected to do certain 'service tasks'. It's important work. We call it 'service work' because because it's really a service to the science, but it's not glorious, it's not glamorous, it's not credited enough. And in the end, when you're looking for academic jobs, these things are not valued."

Royal Society interview participant

4. Collaborating with non-STEM researchers

Reconciling quantitative and qualitative methods can pose workflow challenges for interdisciplinary collaboration between science, technology, engineering and maths (STEM) and non-STEM researchers like arts, humanities, and social sciences. Furthermore, non-STEM researchers face additional obstacles related to funding opportunities and a lack of 'true interdisciplinary inclusion' which can limit meaningful, sustained collaboration[247].

A STEAM approach, where art is directly included into STEM, can help arts complement STEM research[248]. Interdisciplinary programmes such as Institute for Interdisciplinary Data Science and AI at the University of Birmingham[249] and the College of Integrative Sciences and Arts (CISA) at Arizona State University[250] can address this gap by supporting researchers to work across disciplines[251].

Collaboration between AI experts and researchers in non-STEM fields can offer opportunities including[252]:

- Adopting AI-driven methods in non-STEM fields (eg Generative AI in art and photography)[253];
- Leveraging artists' creativity and expertise for AI-based research (eg user experience design)[254];
- Including social science and humanities perspectives for responsible AI, AI safety, and research ethics discussions[255].
- Adopting non-STEM participatory research methods, including citizen science and open dialogues, to enhance trust, transparency, and inclusivity in AI-based research[256].

247 The Royal Society interviews with scientists and researchers. 2022 - 2023

248 The Royal Society. 2023 Science in the metaverse: policy implications of immersive technologies. See https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/ (accessed 21 December 2023).

249 University of Birmingham. The Institute for Interdisciplinarity Data Science and AI. See https://www.birmingham.ac.uk/research/data-science/index.aspx (accessed 11 January 2024)

250 Arizona State University. See: https://news.asu.edu/20230407-university-news-asu-college-integrative-sciences-arts-reorganizes-3-new-schools. (accessed 13 December 2023)

251 The Royal Society. 2023 Science in the metaverse: policy implications of immersive technologies. See https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/ (accessed 21 December 2023).

252 The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023.

253 Wu T, Zhang, SH. 2024 Applications and Implication of Generative AI in Non-STEM Disciplines in Higher Education. *In:* Zhao, F., Miao, D. (eds) AI-generated Content. AIGC 2023. Communications in Computer and Information Science, vol 1946. Springer, Singapore. (doi.org/10.1007/978-981-99-7587-7_29)

254 The Royal Society. 2023 Science in the metaverse: policy implications of immersive technologies. See https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/ (accessed 21 December 2023).

255 Zeller F, Dwyer L. 2022 Systems of collaboration: challenges and solutions for interdisciplinary research in AI and social robotics. *Discover Artificial Intelligence*, **2**. 12. (https://doi.org/10.1007/s44163-022-00027-3)

256 UNESCO Recommendation on Open Science. 2021. See: https://www.unesco.org/en/legal-affairs/recommendation-open-science (accessed 6 February 2024)

Emerging research skills in AI-based research

The increased application of AI highlights the need for foundational AI and data skills across research fields[257]. While advanced data and programming expertise may not always be required, nor will AI methods be necessary for all research areas, awareness of the latest tools and techniques can increase accessibility and benefits for science. The following areas of emerging research skills were drawn from Royal Society interviews and roundtables involving scientists from across academia and industry.

1. **Specialist data skills**
   Demand for specialised data skills has increased as more research fields adopt AI. Data scientists and engineers are predicted to have the fastest growth between 2023 and 2027[258] and roles including data stewards and reproducibility experts are gaining increasing value as AI challenges emerge (See Chapter 2). However, critical data management tasks including curation, cleaning, and quality assurance are often undervalued as "service work"[259]. While enhancing data literacy across disciplines can aid AI uptake, the rapid advancement of AI tools risks outpacing training on effective data practices.

2. **AI literacy training**
   The skills gap can also include a lack of understanding of challenges related to bias, reproducibility, and data requirements when employing AI models. Upskilling scientific domain experts is essential for envisioning innovative AI applications. Organisations like Cambridge Spark provide educational resources for data science, addressing the skills gap through apprenticeships, corporate training, or skills bootcamps[260]. Additionally, the EU is fostering education and skills through various initiatives under Horizon Europe, such as the European Institute for Innovation and Technology Knowledge and Innovation Communities (EIT KICs), the European Innovation Council, and the Erasmus+ programme[261].

> "As an ageing scientist who is not a bioinformatics person, I find a lot of these things quite impenetrable. Often you don't have the confidence to find something and you're going to need other colleagues to really go in there and have an in-depth look."
>
> **Royal Society interview participant**

257 The Royal Society. 2019. Dynamics of data science skills: How can all sectors benefit from data science talent. See https://royalsociety.org/-/media/policy/projects/dynamics-of-data-science/dynamics-of-data-science-skills-report.pdf (accessed 6 January 2024)

258 World Economic Forum. 2023 The Future of Jobs Report 2023. See https://www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf (accessed 30 January 2024)

259 The Royal Society roundtable on reproducibility, April 2023

260 Cambridge Spark. See https://www.cambridgespark.com/ (accessed 1 August 2023)

261 Petkova, D, Roman, L. 2023 AI in science: Harnessing the power of AI to accelerate discovery and foster innovation – *Policy brief, Publications Office of the European Commission, Directorate-General for Research and Innovation.* (doi/10.2777/401605)

"I have always used machine learning technologies, but I hadn't thought about bad uses of AI. I never had an ethics lecture in my life as a scientist and that was in the last 30 years or so of my educational life. That goes to show that we've messed up over the last 30 years."

Royal Society roundtable participant

3. AI ethics training

Researchers are adopting new AI techniques, such as generative AI, without a full understanding of the ethical implications. This is due to a lack of evidence on potential risks and limited AI ethics training. (See Chapter 5)[262].

Collaboration with ethicists and AI ethics training can help bridge this gap. This is being explored by organisations such as the Montreal AI Ethics Institute, an international non-profit organisation, which aims to 'democratise AI ethics literacy' by providing accessible resources, including a living dictionary, and an AI ethics briefing[263].

Involving researchers in AI assurance activities can also contribute towards building skills to identify vulnerabilities and risks in AI systems. Examples include red teaming (See Box 3) or training in principles like adversarial machine learning[264].

4. Human-in-the-loop systems and skills

AI can augment researchers' skills, assist tasks, or automate processes[265] (See Chapter 1). It can also play a role in supporting human judgement and creativity in scientific endeavours. Defining complementary roles for humans and AI to support scientific research and reskilling for automation (as seen in total laboratory automation case studies[266]) will be necessary. This transition also highlights the need for human-in-the-loop systems for quality control, quality assurance, and adaptation to changing workflow dynamics[267].

262 Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (https://doi.org/10.48550/arXiv.2302.04844)

263 Montreal AI Ethics Institute. See https://montrealethics.ai/. (accessed 26 February 2024.)

264 The Royal Society. 2024. Red teaming large language models (LLMs) for resilience to scientific disinformation. See https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/

265 OECD. 2023. Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research, *OECD Publishing, Paris* (https://doi.org/10.1787/a8d820bd-en).

266 Archetti C, Montanelli A, Finazzi D, Caimi L, Garrafa E. Clinical laboratory automation: a case study. *J Public Health Res*. 2017;6(1):881. (doi: 10.4081/jphr.2017.881)

267 Al Naam YA *et al* 2022 The Impact of Total Automaton on the Clinical Laboratory Workforce: A Case Study. J *Health Leadersh.* **9;**14:55-62. (doi:10.2147/JHL.S362614)

BOX 3

Insights from the Royal Society and Humane Intelligence red-teaming exercise on AI-generated disinformation content[268]

Red teaming refers to the process of actively identifying potential weaknesses, failure modes, biases, or other limitations in a model, technology, or process by having groups 'attack' it.

In the run-up to the UK's 2023 Global AI Safety Summit, the Royal Society and Humane Intelligence brought together 40 postgraduate students in health and climate sciences to scrutinise how potential vulnerabilities in LLMs (Meta's Llama 2) could enable the production of scientific misinformation.

By assuming different 'misinformation actor' roles, participants tested the model's guardrails related to topics of infectious diseases and climate change. In under two hours, they exposed concerning vulnerabilities, including the model's inability to convey scientific uncertainty and its reliance on questionable or ficticious sources.

While guardrails prevented some common disinformation trends, such as those related to COVID-19, participants were still able to generate outputs that distorted verifiable scientific facts arriving at incorrect conclusions.

The exercise demonstrated the value of involving domain experts in AI safety assessments before deployment. Their scientific expertise allowed them to stress test systems in ways that exposed critical failures. Participants also expressed optimism regarding the future of LLM disinformation guardrails and more confidence in using LLMs in their own research. Their insights suggest that red teaming could play a role in enhancing AI literacy within the scientific community.

268 The Royal Society. 2024. Red teaming large language models (LLMs) for resilience to scientific disinformation. See https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/

CASE STUDY 2

AI and material science

Materials science is a field where AI and ML techniques have the potential to be transformative, with wide-ranging societal benefits, provided that appropriate support infrastructure is in place.

The need to develop advanced materials, from new battery materials for energy storage to catalysts to create biodegradable plastics, has been a driver for emerging technologies. Historically, this intricate process relied heavily on a scientist's prior knowledge, intuition, or serendipity to navigate an estimated 10 trillion possible chemistry combinations.

However, AI and ML are now accelerating materials discovery and optimisation[269]. These techniques rapidly screen candidates, predict structures, and offer suggested changes which would have overwhelmed manual approaches[270]. In turn, AI-based workflows can lead to time efficiencies, allowing more ideas to progress from conception to commercialisation within years instead of decades[271].

Materials design and prediction

Modelling and simulation are well-established techniques within materials science. Density Functional Theory (DFT)[272] is one of the most common modelling methods and is an important tool in materials modelling. It allows for accurate calculations of materials behaviour, although it does not work well for certain classes of materials or for certain properties that are important to material behaviour. It is also a computationally expensive method and, as such, is limited to materials of low to medium complexity[273]. Other techniques such as Monte-Carlo simulation and molecular dynamics are also commonly used but are similarly computationally expensive for complex materials[274].

AI and ML techniques can be used to predict the structure and properties of materials. An example of this is using generative algorithms and foundation models to predict what materials might exhibit desirable properties[275]. However, the current materials knowledge base is large and disparate, with data often being incomplete, noisy, inconsistently formatted, and poorly labelled[276]. Significant amounts of materials data are sequestered in journals without open-access, or never published at all. Furthermore, data from negative or unsuccessful experiments are not routinely published.

269 Davies *et. al.* 2016 Computational screening of all stoichiometric inorganic materials. Chem. **1**, 617-627. (https://doi.org/10.1016/j.chempr.2016.09.010).

270 Pyzer-Knapp *et. al.* 2023 Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Computational Materials. **8**, 84. (https://doi.org/10.1038/s41524-022-00765-z).

271 Materials Genome Initiative. About the Materials Genome Initiative. See https://www.mgi.gov/about (accessed 14 July 2023).

272 Argaman N, Makov G. 2000 Density functional theory: An introduction. American Journal of Physics. **68**, 69-79. (https://doi.org/10.1119/1.19375).

273 Alberi *et. al.* 2019 The 2019 materials by design roadmap. Journal of Physics D: Applied Physics. **52**, 013001. (https://doi.org/10.1088/1361-6463/aad926).

274 Tao Q, Xu P, Li M, Lu W. 2021. Machine learning for perovskite materials design and discovery. npj Computational Materials. 7, 23. (https://doi.org/10.1038/s41524-021-00495-8).

275 Ross *et. al.* 2022 Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence. **4**, 1256-1264. (https://doi.org/10.1038/s42256-022-00580-7).

276 Liu Y, Zhao T, Ju W, Shi S. 2017 Materials discovery and design using machine learning. Journal of Materiomics. **3**, 159-177. (https://doi.org/10.1016/j.jmat.2017.08.002).

In recent years, there have been several materials databases developed with the goal of aggregating data in consistent formats which can then be used for further research. Examples include the Materials Project[277] and Aflow[278] databases (which both contain computed properties) and the Inorganic Crystal Structure Database (ICSD)[279] and the High Throughput Experimental Materials (HTEM)[280] database (which are both examples of experimental databases). There are also tools to help with the creation and analysis of materials datasets, such as NOMAD[281], ChemML[282], and atomate[283]. These datasets which can be significant in size (eg the Materials Project database currently contains data for more than 150,000 materials), have been facilitating the use of ML for materials discovery.

There have been several success stories in recent years of ML being used for materials discovery, some examples of which are listed in Table 2. A variety of ML and AI techniques, including generative AI, have been used to identify materials with desired properties for a wide range of applications. These have been integrated with established techniques such as DFT, stability calculations and experiments to narrow down the predicted materials[284]. Sustainability of proposed materials could also be used as an objective for predictive models[285], to prevent new, more complex materials being harder to recycle or dispose of safely.

277 Jain *et. al.* 2013 The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials. 1, 011002. (https://doi.org/10.1063/1.4812323).

278 Curtarolo *et. al.* 2012 AFLOWLIB.ORG: A distributed materials properties repository from high-throughput *ab-initio* calculations. Computational Materials Science. 58, 227-235. (https://doi.org/10.1016/j.commatsci.2012.02.002).

279 Physical Sciences Data-Science Service. ICSD. See https://www.psds.ac.uk/icsd (accessed 14 July 2023).

280 Zakutayev *et. al.* 2018 An open experimental database for exploring inorganic materials. Scientific Data. 5, 180053. (https://doi.org/10.1038/sdata.2018.53).

281 Draxl C, Scheffler M. 2019 The NOMAD laboratory: from data sharing to artificial intelligence. Journal of Physics: Materials. 2, 036001. (https://doi.org/10.1088/2515-7639/ab13bb).

282 ChemML. See https://hachmannlab.github.io/chemml/ (accessed 14 July 2023).

283 Atomate. See https://atomate.org/ (accessed 14 July 2023).

284 DeCost *et. al.* 2020 Scientific AI in materials science: a path to a sustainable and scalable paradigm. *Machine Learning: Science and Technology*. **1**, 033001. (https://doi.org/10.1088/2632-2153/ab9a20).

285 Raabe D, Mianroodi J, Neugebauer J. 2023 Accelerating the design of compositionally complex materials via physics-informed artificial intelligence. *Nature Computational Science*. **3**, 198-209. (https://doi.org/10.1038/s43588-023-00412-7)

TABLE 2

Examples of machine learning in materials discovery

Researchers	Result
Lyngby et. al.[286]	Predicted 11,630 new, stable 2D materials.
Yao et. al.[287]	Found 2 new 'invar alloys' which have a low thermal expansion and can be useful for several applications.
Vasylenko et. al.[288]	Identified 4 new materials, including materials that have desirable properties for use in solid state batteries.
Sun et. al.[289]	An approach for pre-screening for new organic photovoltaic materials.
Stanev et. al.[290]	Identified >30 potential high-temperature superconducting materials.

286 Lyngby P, Sommer Thygesen K. 2022. Data-driven discovery of 2D materials by deep generative models. npj Computational Materials. 8, 232. (https://doi.org/10.1038/s41524-022-00923-3).

287 Rao *et. al.* 2022 Machine learning-enabled high-entropy alloy discovery. *Science*. **378**, 78-85. (https://doi.org/10.1126/science.abo4940).

288 Vasylenko *et. al.* 2021 Element selection for crystalline inorganic solid discovery guided by unsupervised machine learning of experimentally explored chemistry. *Nature Communications*. **12**, 5561. (https://doi.org/10.1038/s41467-021-25343-7).

289 Sun *et. al.* 2019. Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. Science Advances. 5, 11. (https://doi.org/10.1126/sciadv.aay4275).

290 Stanev *et. al.* 2018. Machine learning modelling of superconducting critical temperature. npj Computational Materials. **4**, 29. (https://doi.org/10.1038/s41524-018-0085-8).

Automated experimentation

Another important application of AI in materials is for automated experimentation (AE). The central idea is to integrate AI and robotics in a closed experimental loop, whereby an AE system can undergo an iterative experimental process based on prior knowledge, improving the output after each iteration[291]. There is human intervention to initialise the experiment and define the objectives.

Increased adoption of AE systems in labs would have a significant impact on materials design and experimentation. Firstly, it would increase the speed of research at a reduced per-experiment cost[292]. This would free up scientists to work on other tasks, rather than focusing on routine, time-consuming experiments, and would also enable faster development and commercialisation of technologically relevant materials[293]. A global network of integrated AE systems would increase the accessibility of materials science research, opening advanced techniques to research groups with fewer resources.

Although AE technologies exist and are being used for materials research, there are several barriers which need to be addressed prior to wider spread adoption. Firstly, experimental hardware and software improvements would be needed, including non-proprietary interfaces and tools for effective characterisation of samples. Collaborative partnerships between materials scientists and AI experts would be beneficial, as well as an understanding that human involvement will be necessary where there are safety or ethical concerns. Finally, AE systems would need access to a broad range of data, including metadata and negative results, to supply to the system as pre-knowledge, requiring standardisation of data management and sharing.

There are several examples of AE systems being successfully used for materials research. One such example is the 2016 demonstration by Nikolaev *et al*[294]. of the first use of the Autonomous Research System to optimise the growth of carbon nanotubes, a material which has potential uses in carbon capture technologies as well as an astounding array of current applications[295]. However, carbon nanotubes are expensive; the price depends on configuration and quality, but 1 gram retails from around £100 to more than £1200. AE can be used to improve the growth of carbon nanotubes by rapidly iterating growth parameters for property optimisations and greater yields.

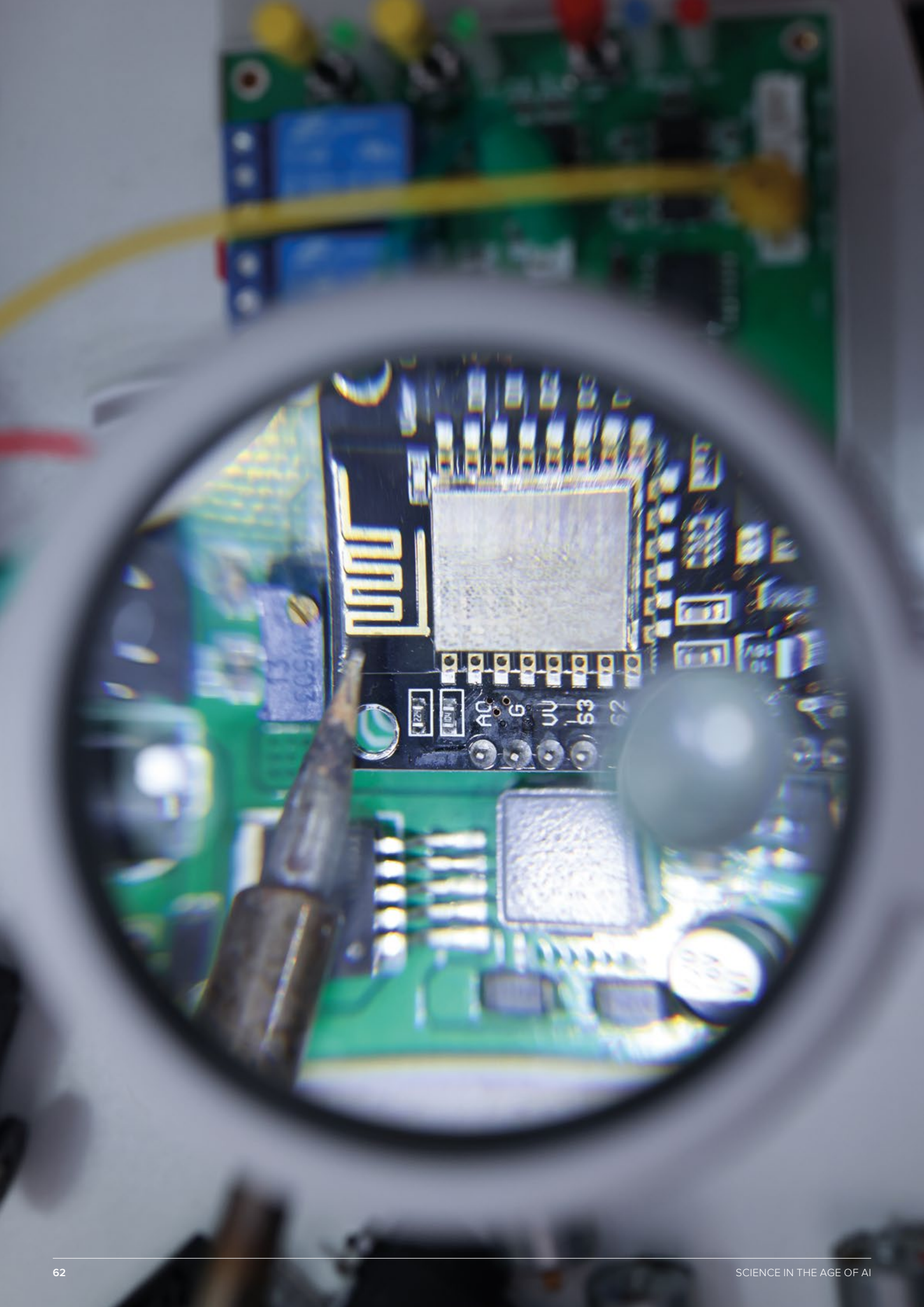
291 Stach *et. al.* 2021 Autonomous experimentation systems for materials development: A community perspective. Matter. **4,** 2702-2726. (https://doi.org/10.1016/j.matt.2021.06.036).

292 Stein H, Gregoire J. 2019 Progress and prospects for accelerating materials science with automated and autonomous workflows. Chemical Science. 10, 9640. (https://doi.org/10.1039%2Fc9sc03766g)

293 Maruyama *et. al.* 2023 Artificial intelligence for materials research at extremes. MRS Bulletin. **47**, 1154-1164. (https://doi.org/10.1557/s43577-022-00466-4).

294 Nikolaev *et. al.* 2016 Autonomy in materials research: a case study in carbon nanotube growth. npj Computational Materials. 2, 16031. (https://doi.org/10.1038/npjcompumats.2016.31).

295 De Volder M, Tawfick S, Baughman R, Hart J. 2013 Carbon Nanotubes: Present and Future Commercial Applications. Science. 339, 535-539. (https://doi.org/10.1126/science.1222453).



Chapter four
Research, innovation and the private sector

Left
Electronic circuits.
© iStock / onuma Inthapong.

Research, innovation and the private sector

The large investment in AI by the private sector and its significance in scientific research present various implications. These include the centralisation of critical digital infrastructure[292]; the attraction of talent away from academia to the private sector[293]; and challenges to open science[294].

The influence of the private sector in the development of AI for science is not extraordinary. Historically, the automation of tasks has been driven by industry actors in the pursuit of reduced labour costs and greater scalability[295]. Today, the private sector continues to play a prominent role in advancing scientific research, with many companies having AI-driven scientific programmes such as Alphabet's Google DeepMind and Microsoft's AI for Science[296].

The role of the private sector in science is also expanding as many companies contribute to provisioning essential resources like computational power, data access and novel AI technologies to the wider research community[297].

This chapter examines the growing role of the private sector in science, drawing on a commissioned review of the global AI patent landscape, which describes the distribution of ownership, development and impact of AI technologies. It also gathers perspectives from a horizon-scanning workshop on AI safety risks and a commissioned historical review.

292 Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

293 Gofman M, Jin Z. 2022 Artificial Intelligence, Education, and Entrepreneurship. *Journal of Finance, Forthcoming*. (https://doi.org/10.1111/jofi.13302)

294 *Ibid.*

295 Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

296 Microsoft. Microsoft Research. AI4Science. See https://www.microsoft.com/en-us/research/lab/microsoft-research-ai4science/ (accessed 21 December 2023)

297 Kak A, Myers West S, Whittaker M. 2023 Opinion: Make no mistake – AI is owned by Big Tech. *MIT Technology Review.* See https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/. (accessed 21 December 2023)

The changing landscape of AI technologies in scientific research

Public and private sector investment in AI for scientific advancement is increasing[298]. For instance, the UK Government, through UKRI, committed £54 million to universities across the country to support responsible AI development and fund AI-based research projects and a further £50 million to accelerate research ventures with industry and the third sector[299]. However, in 2023, Microsoft invested over £8 billion into OpenAI[300] and Meta pledged around £26.5 billion in expanding their 'AI capacity'[301]. It is estimated that, in 2022, the private sector accounts for 67% of AI investment in the EU, with the public sector contributing 33%[302].

One method of understanding the changing landscape of AI technologies in scientific research is by looking at intellectual property (IP) trends. IP refers to creations of the mind, such as inventions, designs, or literacy and artistic work. IP law includes copyright, trademark, trade secrets and patents which grant exclusive rights to inventors or assignees for a limited time in exchange for public disclosure of the invention[303]. This section primarily draws on patent trends (See Box 4) and addresses the changing landscape of AI technologies in scientific research.

298 IP Pragmatics, 2024 Artificial intelligence related inventions. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

299 UK Government. £54 million boost to develop secure and trustworthy AI research. Gov.UK. See https://www.gov.uk/government/news/54-million-boost-to-develop-secure-and-trustworthy-ai-research (accessed 21 December 2023)

300 Bass D. 2023 Microsoft invests $10 billion in ChatGPT maker OpenAI. *Bloomberg*. 23 January 2023. See https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai (accessed 21 December 2023).

301 Targett E. 2023 Meta to spend up to $33 billion on AI, as Zuckerberg pledges open approach to LLMs. *The Stack*. 27 April 2023. See https://www.thestack.technology/meta-ai-investment/ (accessed 21 December 2023).

302 *Ibid.*

303 Intellectual property and your work. Gov.UK. See: https://www.gov.uk/intellectual-property-an-overview (accessed March 22 2024)

BOX 4

IP Pragmatics 2023: Global patent landscape analysis

The Royal Society commissioned a global patent landscape review on AI technology patents for scientific research. The analysis assessed the ownership, development, and impact of AI patents among countries, organisations, and industries in the past 10 years. It also identified key players, trends, and potential implications for the scientific community.

This analysis defined AI as the study in computer science aimed at developing machines and systems that can carry out tasks considered to require human intelligence, such as ML or ANNs. Key search terms and relevant International Patent Classification (IPC) codes were used to identify AI-related patents comprehensively (see IP Pragmatics report for more information)[304].

While patent landscape reviews can provide useful findings, there are limitations. Obtaining complete global patent data can be challenging, and the analysis may present an incomplete picture due to limitations in data availability or accessibility. While delay times may vary between jurisdictions by a few months, there is an 18-month delay between priority application and patent publication in most of the main global territories. This means that data in this review from 2021 – 2023 is incomplete[305].

Additionally, patent data alone does not capture the full extent of AI research and development, as some innovations may not be patented or may be protected through other IP laws including copyright, trade secrets and trademarks.

304 IP Pragmatics, 2024 Artificial intelligence related inventions. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

305 *Ibid.*

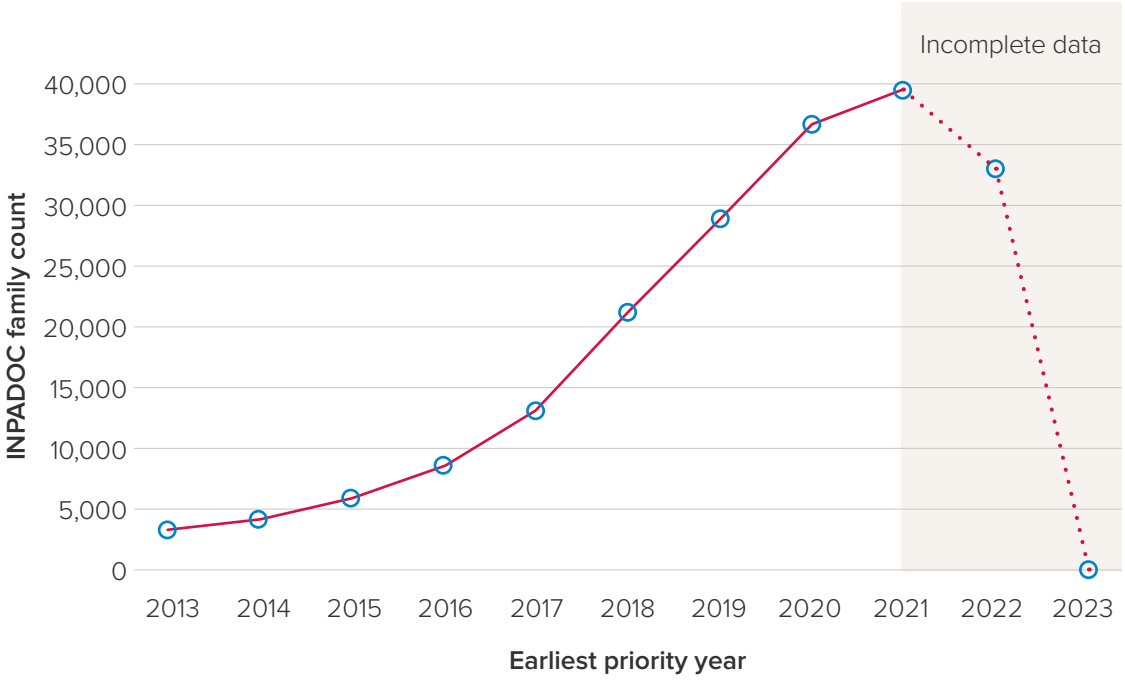
1. Growth in the AI patent landscape

The AI patent landscape has surged in the past decade, with approximately 74% of the total patent filings occurring in the last five years[306] (see Figure 2). In 2022, the market value reached £109.718 billion, with a projected compound annual growth rate (CAGR) of 37.3% from 2023 to 2030[307].

Notably, China leads in AI patent filings, holding approximately 62% of the landscape, followed by the United States with around 13.2%[308]. The Asia-Pacific region is projected to have the highest CAGR (48.6%) between 2022 and 2027, in comparison to the US (43%) and Europe (46.5%). This finding complements the reported increase in AI innovation in the healthcare industry in China and India[309,310].

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).

306 *Ibid.*

307 Grand View Research. See https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market# (Accessed 21 December 2023)

308 IP Pragmatics, 2024 Artificial intelligence related inventions. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

309 Nair M, Sethumadhaven A. 2022 AI in healthcare: India's trillion-dollar opportunity. *World Economic Forum.* See https://www.weforum.org/agenda/2022/10/ai-in-healthcare-india-trillion-dollar (accessed 21 December 2023)

310 Olcott E. 2022, China sets the pace in adoption of AI in healthcare technology. *Financial Times*. 31 January 2022. See: https://www.ft.com/content/c1fe6fbf-8a87-4328-9e75-816009a07a59 (accessed 21 December 2023)

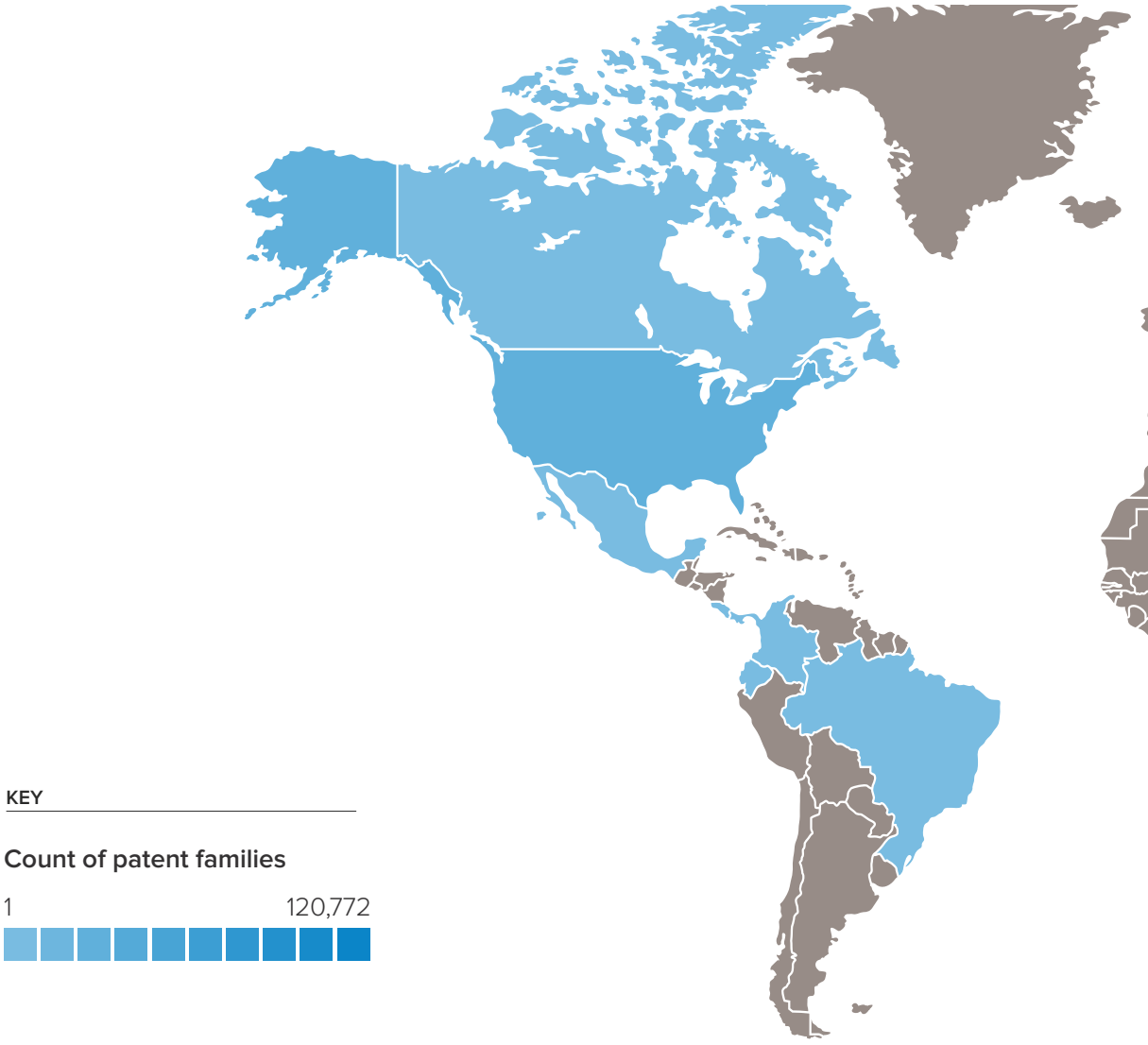
2. Global market shares in AI for science

The use of AI in the science and engineering market is being driven primarily by the demand for AI technology to drive innovation and economic growth. As such, there is a correlation between patent filing trends and global market shares[311].

North America, with its rich concentration of technology firms and skilled professionals, dominates this market[312]. In Europe, Germany leads, but the United Kingdom stands out with a significant 14.7% share in the AI for life sciences market and has the region's highest forecasted CAGR of 47.9%[313].

FIGURE 3

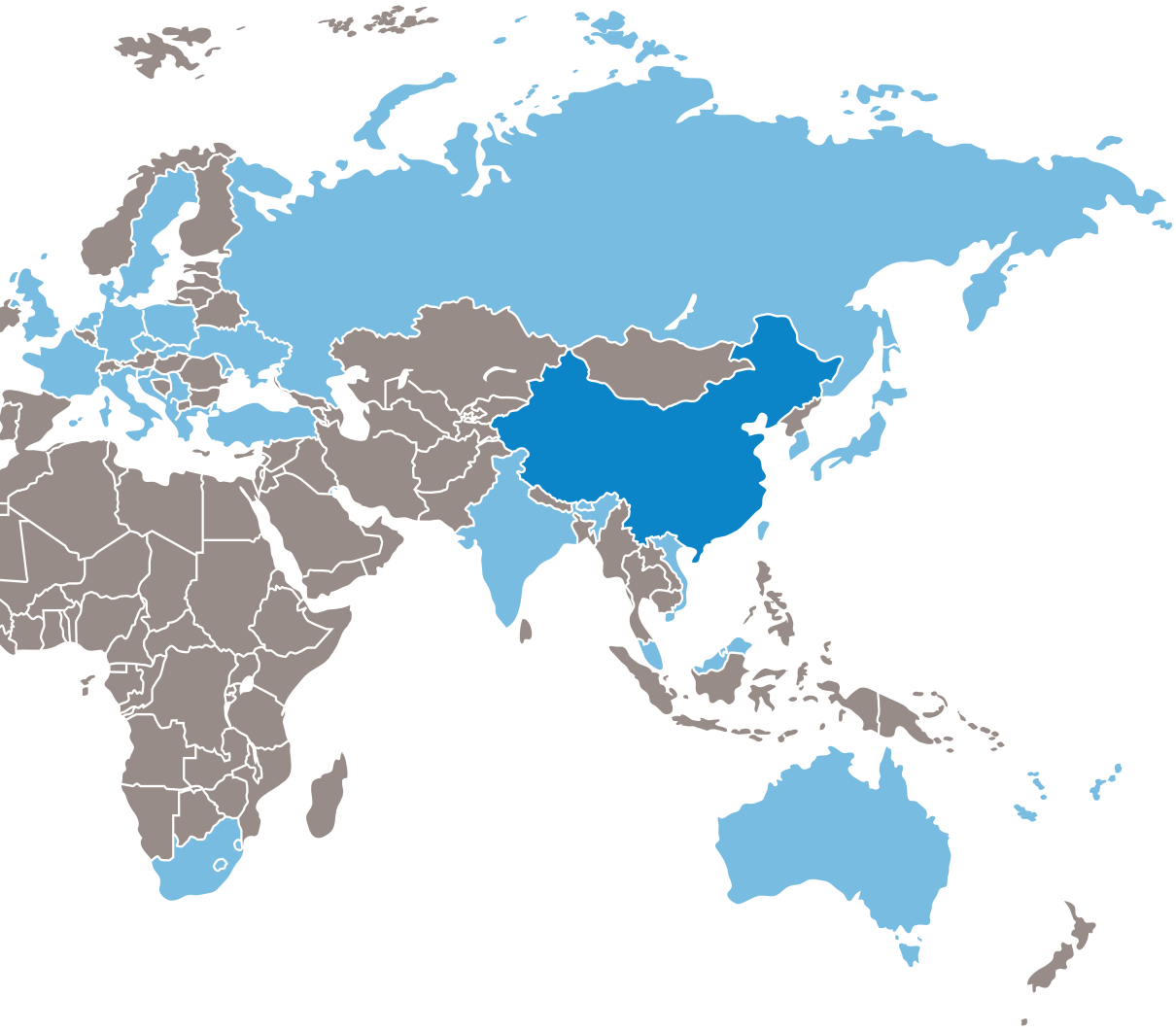
Global distribution of the number of AI-related patent families by 1st priority country



311 BBC Research. 2022 Global Markets for Machine Learning in the Life Sciences. October 2022. See https://www.bccresearch.com/market-research/healthcare/global-markets-for-machine-learning-in-life-sciences.html (accessed 21 December 2023)

312 *Ibid.*

313 *Ibid.*

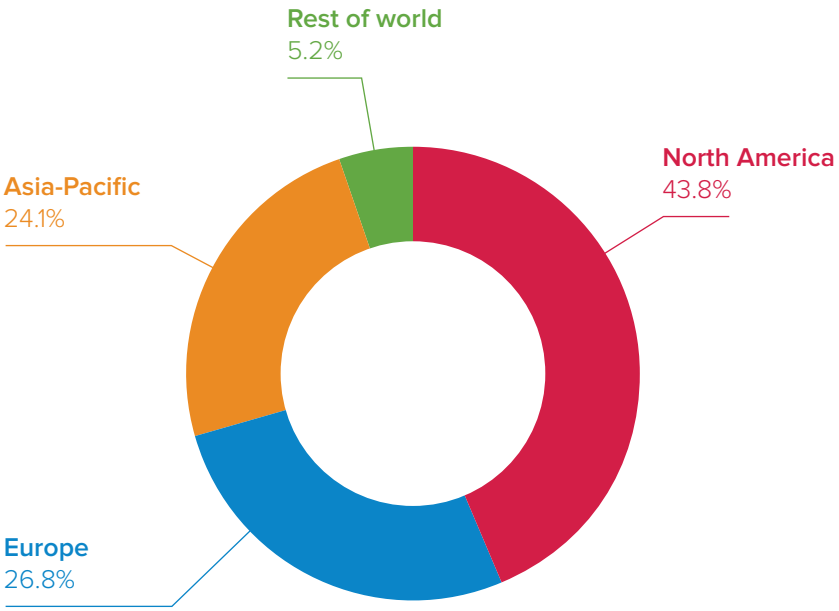


The UK, ranking 10th globally and 2nd in Europe for patent filings, demonstrates strong growth potential[314]. The UK Intellectual Property Office (UKIPO) adopts a more patentee-friendly approach to examining computer-implemented and AI inventions compared to the European Patent Office (EPO), as underscored by recent decisions like Emotional Perception AI vs Comptroller General of Patents[315].

This case led to an adjustment in UKIPO's examination practices, removing specific guidance on ANNs[316]. While this decision has been appealed and currently awaits review, recent rulings have reinforced the UK's position as a preferred region for AI-related IP protection, bolstering its role as a key player in AI innovation.

FIGURE 4

Global market shares of machine learning in the life sciences, by region, 2021 (%)



Source: BCC Research.

314 IP Pragmatics. 2024 Artificial intelligence related inventions. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

315 Emotional perception ai ltd v comptroller-general of patents, designs and trademarks. 2023. Find case law – The National Archives. See https://caselaw.nationalarchives.gov.uk/ewhc/ch/2023/29482948 (accessed 4 March 2024).

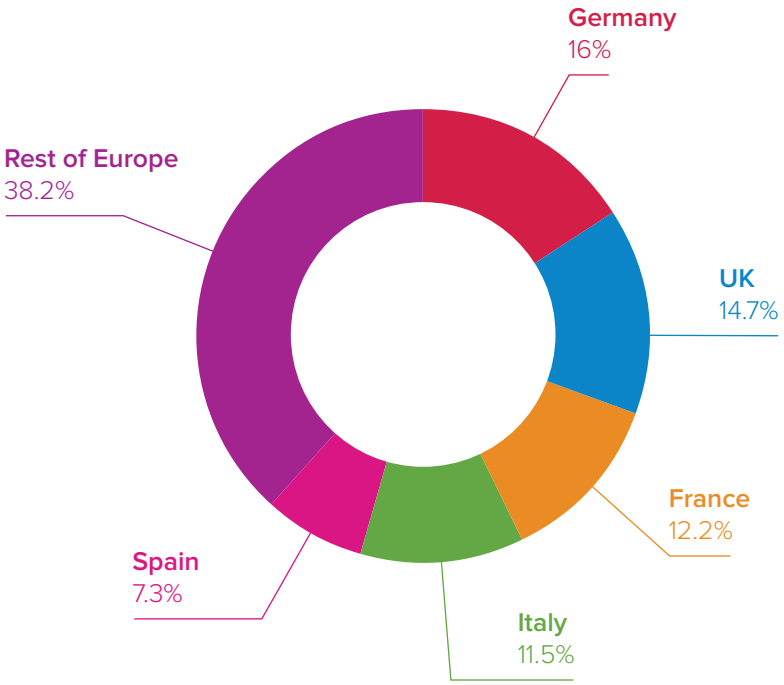
316 Examination of patent applications involving Artificial Neural Networks (ANN). Gov.UK. See https://www.gov.uk/government/publications/examination-of-patent-applications-involving-artificial-neural-networks/examination-of-patent-applications-involving-artificial-neural-networks-ann (accessed 4 March 2024).

However, the global landscape is marred by disparities. The costly and intricate patent application processes, particularly in regions like Africa, pose considerable barriers. For example, patenting in Africa through the African Regional Intellectual Property Organisation costs over £29,000, significantly higher than in the UK, priced at around £1,900.

Despite a surge in African technology hubs, high IP registration expenses and lack of a unified system hamper patenting[317]. Initiatives like the Pan-African Intellectual Property Organisation aim to address these challenges, although they currently face operational delays[318].

FIGURE 5

European market shares of machine learning in the life sciences, by country, 2021 (%)



Source: BCC Research.

317 Lewis J, Schneegans S, Straza T. 2021 UNESCO Science Report: The race against time for smarter *development*. *UNESCO Publishing*. See: https://unesdoc.unesco.org/ark:/48223/pf0000377250 (accessed 22 March 2024)

318 *Ibid.*

3. **Key players in the AI for Science patent landscape**
   In terms of technological impact (indicated by the number of times that a patent is cited by a later patent or forward citations) the US stands out for having valuable patents. Comparatively, despite India's significant growth in AI patent filings, it has not yet achieved large technological impact. While the UK, though representing a smaller portion of the patent landscape, demonstrates research and innovation influence, ranking among the highest globally[319].

   The analysis of the top 20 assignees in AI-related patents underscores the active involvement of both industry and academic entities within the broader scientific and engineering research sphere. Notably, companies such as Canon, Alphabet, Siemens, IBM, and Samsung have emerged as key contributors, with substantial patent portfolios that wield considerable influence across scientific and engineering domains. Despite the dominance of commercial entities in most regions, academic institutions including the University of Oxford, Imperial College London, and University of Cambridge feature prominently among the top patent filers in the UK [320], suggesting blend of academic-industry collaboration and independent contributions[321].

Challenges related to the role of the private sector in AI-based science

In addition to looking at patenting trends, the Royal Society explored the challenges of private sector involvement in AI-based scientific research. Ahead of the Global AI Safety Summit hosted by the United Kingdom in 2023, the Royal Society and the UK's Department for Science, Innovation and Technology (DSIT) convened a horizon scanning workshop on the safety risks of AI in scientific research[322]. Challenges identified include:

1. **Private sector dominance and centralisation of AI-based science development**
   Centralisation of AI development under large technology firms (eg Google, Microsoft, Amazon, Meta and Alibaba) could lead to corporate dominance over infrastructure critical for scientific progress. This includes ownership over massive datasets for training AI models, vast computing infrastructures, and top AI talent[323].

   Centralisation can limit wider participation in steering the AI research agenda and may result in a small number of decision-makers to shape what research is conducted and published from influential industrial labs. For instance, the high-profile and controversial dismissal of AI researcher Dr Timnit Gebru from Google highlighted the opaque internal decision-making in private sector research units.

319 IP Pragmatics. 2024 Artificial intelligence related inventions. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

320 *Ibid.*

321 Legislation.Gov.UK. Copyright, Designs and Patents Act 1988. See: https://www.legislation.gov.uk/ukpga/1988/48/contents

322 Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See https://royalsociety.org/current-topics/ai-data/ (accessed 7 May 2024)

323 Kak A, Myers West S, Whittaker M. 2023 Opinion: Make no mistake – AI is owned by Big Tech. MIT Technology Review. 5 December 2023 See https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/. (accessed 21 December 2023)

This is an example of a misalignment between corporate interest to protect their market advantage and academic values to openly scrutinise the societal impact of advancing technology[324].

2. **Overreliance on industry-driven tools and benchmarks for AI-based science**
Industry's growing influence in setting key benchmarks[325], developing cutting-edge models[326], and steering academic publications[327] is centralising control over AI ecosystems, encompassing hardware, software, and data[328].

The patent system wields substantial influence over markets, stimulating competition and growth in all sectors. Through this system, patent proprietors can ensure investment return and control of technology dissemination. However, without oversight, there is a risk that these entities will prioritise commercial interests over broader scientific advancement, controlling critical infrastructure, databases, and algorithms to maintain their market dominance[329].

The concentration of power and resources could not only hinder competition and transparency but also establish single points of failure, raising concerns about the resilience and openness of AI-based scientific research[330].

324 Hao K. 2020 We read the paper that forced Timnit Gebru out of Google. Here's what it says. *MIT Technology Review*. 4 December 2020. See https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/ (accessed 12 July 2023)

325 Hodak, M., Ellison, D., & Dholakia, A. (2020, August). Benchmarking AI inference: where we are in 2020. In *Technology Conference on Performance Evaluation and Benchmarking* (pp. 93-102). Cham: Springer International Publishing.

326 *Ibid.*

327 Ahmed N, Wahed M, Thompson, N. C. 2023. The growing influence of industry in AI research. Science, 379(6635), 884-886. (https://doi: 10.1126/science.ade2420)

328 The Royal Society interviews with scientists and researchers. 2022 - 2023

329 Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

330 Westgarth, T., Chen, W., Hay, G., & Heath, R. 2022 Understanding UK Artificial Intelligence R&D commercialisation and the role of standards. See https://oxfordinsights.com/wp-content/uploads/2023/10/DCMS_and_OAI_-_Understanding_UK_Artificial_Intelligence_R_D_commercialisation__accessible-1.pdf (accessed 21 December 2023)

BOX 5

The role of the private sector in patenting medicine and pharmaceutical inventions

An analysis of AI patents in medicine and pharmaceutical interventions shows that while Harvard University and Massachusetts Institute of Technology (MIT) were pioneers in this area, patent portfolios held by Roche and IBM appear to be most valuable in this sector[331]. Alphabet has expanded its influence through subsidiaries like Google DeepMind, which developed AlphaFold – an AI system that has revolutionised protein structure prediction[332]. This innovation marks a significant shift in the medical patent landscape[333,334], prompting other technology giants like Microsoft, to invest in similar technologies. This highlights the commercial potential and competitive dynamics in this sector[335].

However, Alphabet's strategy of filing cluster of patents, considered 'patent ring-fencing', suggests a broader trend of firms leveraging their IP to safeguard and expand their market position[336]. Such approaches not only protect against infringement but also prevent competitors from developing adjacent technologies, reinforcing Alphabet's – and by extension, Google DeepMind's – dominance in AI-driven medical research.

331 IP Pragmatics. 2024 Artificial intelligence related inventions. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

332 Google DeepMind. Technology: AlphaFold. See https://deepmind.google/technologies/alphafold/ (accessed 21 December 2023)

333 Borkakoti N, Thornton J.M., 2023. AlphaFold2 protein structure prediction: Implications for drug discovery. Current opinion in structural biology, **78**, p.102526 (https://doi.org/10.1016/j.sbi.2022.102526)

334 Jumper, J., *et al*. 2021 Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), pp.583-589. (doi: 10.1038/s41586-021-03819-2)

335 IP Pragmatics. 2024 Artificial intelligence related inventions. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

336 *Ibid.*

3. The private sector and open science

The commercial incentives driving private ownership of data for AI-based research could restrict open science practices. This limits non-industry scientists' ability to equitably contribute to and scrutinise data for AI systems alongside industry counterparts.

Privately held data is often commercially sensitive and could necessitate non-disclosure agreements, potentially affecting research integrity. Data considered low risk initially may later gain commercial value and get withdrawn, as seen with some social media companies tightening data access following the surge of LLMs training on public data[337,338].

Alternative monetisation approaches like encouraging the licensing of data lakes and utilising database provisions can provide a more open and pragmatic approach to data sharing[339].

Further approaches include changes to legislation such as the requirements for social media companies to share data in the European Digital Services Act[340] and the principles for intervention to unlock the value of data across the economy in the UK's National Data Strategy[341]. Additionally, technical approaches include privacy enhancing technologies[342] and cyber-security legislation to provide legal measures and ensure safer hardware and software[343].

Open-source code and platforms do also offer some advantages to private sector organisations, including speed and cost-effectiveness, but also have significant limitations including lack of support, security risks, and compatibility. For example, industrial partnerships for mutual benefits, such as the partnership between Siemens and Microsoft, can drive cross-industry AI adoption by sharing software, hardware and talent[344]. During the COVID-19 pandemic, some private organisations relinquished patent rights for the common good, with leading technology companies donating their patents to open-source initiatives[345].

337 Isaac M. 2023 Reddit wants to get paid for helping to teach big AI systems. *The New York Times*. 18 April 2023. See https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html (accessed 21 December 2023).

338 Murphy H. 2023 Elon Musk rolls out paywall for Twitter's data. *The Financial Times*. 29 April 2023. See https://www.ft.com/content/574a9f82-580c-4690-be35-37130fba2711 (accessed 21 December 2023).

339 Grossman R L. 2019 Data lakes, clouds, and commons: A review of platforms for analyzing and sharing genomic data. *Trends in Genetics*, *35*(3), pp.223-234. (https://doi.org/10.1016/j.tig.2018.12.006)

340 European Commission. The Digital Services Act. See https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en (accessed 5 February 2024)

341 Department for Science, Innovation and Technology. National Data Strategy. 5 December 2022. See https://www.gov.uk/guidance/national-data-strategy (accessed 5 February 2024)

342 The Royal Society. 2023 From privacy to partnership. See https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/ (accessed 21 December 2023).

343 European Commission. Directive on measures for a high common level of cybersecurity across the Union (NIS2 Directive). See https://digital-strategy.ec.europa.eu/en/policies/nis2-directive (accessed 22 February 2024)

344 Siemens. 2024 Siemens and Microsoft partner to drive cross-industry AI adoption. See https://press.siemens.com/global/en/pressrelease/siemens-and-microsoft-partner-drive-cross-industry-ai-adoption (accessed 26 February 2024)

345 UNESCO Recommendation on Open Science. 2021. See: https://www.unesco.org/en/legal-affairs/recommendation-open-science (accessed 6 February 2024)

4. The private sector's role in AI safety

Private sector dominance in AI for science also poses challenges to AI safety. Organisations and institutions leading AI development often determine their own ability to assess harm, establish safeguards, and safely release their models. As described by OpenAI in the paper behind the release of GPT-4, commercial incentives and safety considerations can come into tension with scientific values such as transparency and open science practices[346].

Hugging Face, an open-source organisation, suggests evaluating the trade-offs for safe and responsible release as illustrated in the Gradient of System Access[347] (see Figure 6). Similar frameworks can be considered and developed by scientific communities to assess the conditions under which releasing training data is safe, allowing them to contribute to scientific progress while reducing potential for harm and misuse.

Universities can also play a crucial role in advancing AI safety, by promoting ethical research standards or incentivising academic research on AI harms. However, they do not have the same capabilities as large technology companies to institute robust safeguards and best practices across all aspects of complex AI development. Recently, national governments have been placing greater significance on AI safety discussions. Since the Global AI Safety Summit in November 2023, the UK has launched the AI Safety Institute[348] while the US announced the US AI Safety Institute under the National Institute of Standards and Technology (NIST)[349].

346 OpenAI *et al*. 2023 Gpt-4 technical report. arxiv 2303.08774. *View in Article*, *2*, p.13. (https://doi.org/10.48550/arXiv.2303.08774)

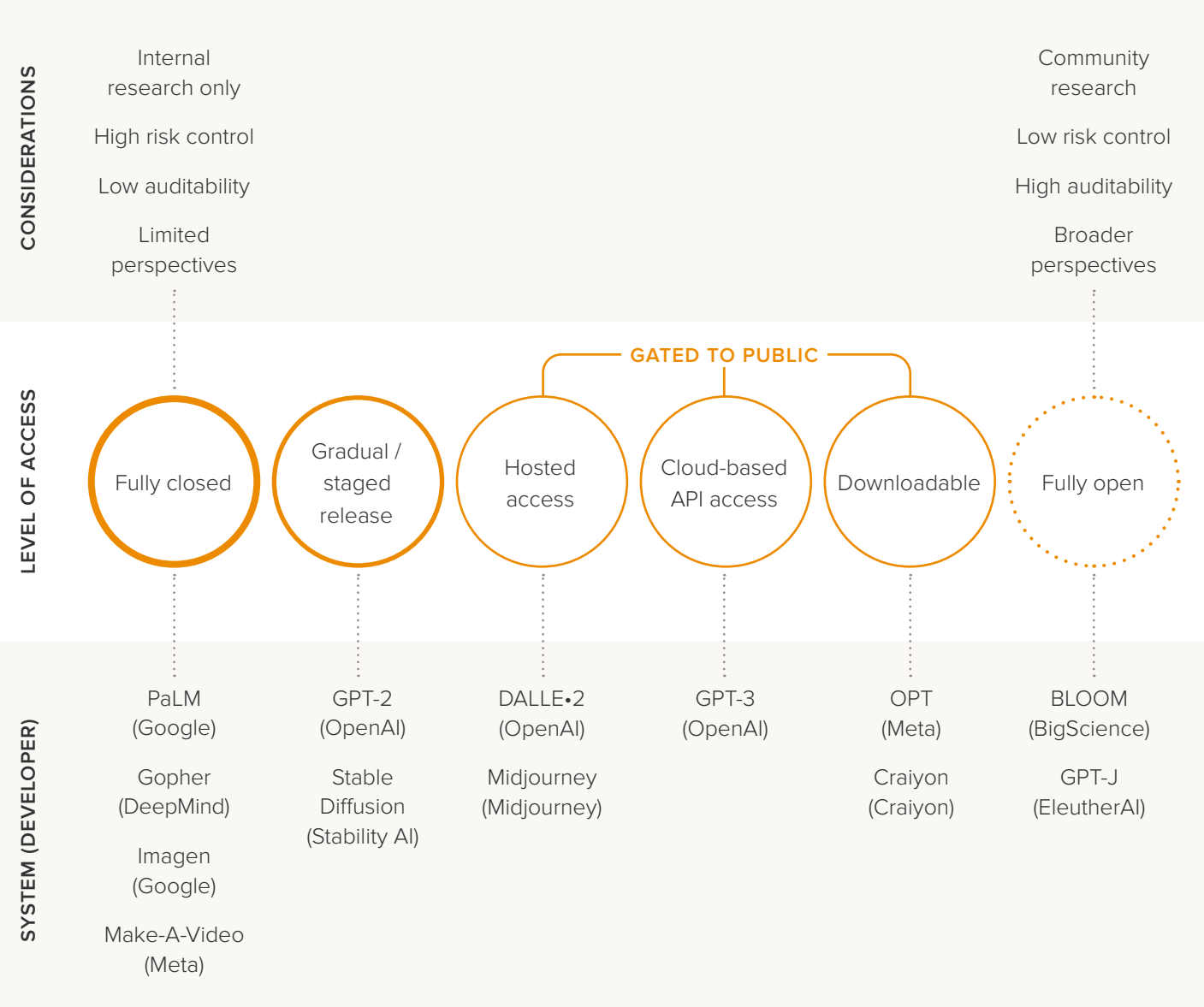
347 Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (https://doi.org/10.48550/arXiv.2302.04844)

348 Gov.UK. 2024 Introducing the AI Safety Institute. See https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute)(accessed 26 February 2024)

349 NIST. 2024 U.S. Artificial Intelligence Safety Institute. See https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute (accessed 26 February 2024)

FIGURE 6

Reproduction of the Gradient of System Access developed by Hugging Face



Source: Hugging Face[350].

350 Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 111-122). (https://doi.org/10.48550/arXiv.2302.04844)

"The freedom, innovation and creativity of academia with the resource and structure and management of the private sector... it's been completely liberating."

Royal Society interview participant referring to joint academic-industry roles

Opportunities for cross-sector collaboration

Cross-sector collaboration offers significant opportunities, leveraging the innovative and educational strengths of academia with the resources and practical focus of industry[351]. Despite concerns about the patent system centralising AI development, it can also foster collaboration. Published patent applications enhance technological transparency and provide a revenue stream that can support joint ventures between universities and industry.

However, the increasing presence of the private sector in AI-based science funding raises concerns that industry's influence might shift the focus from fundamental research to applied science[352]. This shift could exacerbate the 'brain drain'[353], where a significant flow of AI talent leaves academia for the private sector[354], driven by higher salaries, advanced resources and the opportunity to work on practical applications[355].

To counter this trend, initiatives like the UK's Life Sciences Innovative Manufacturing Fund[356] (which includes £17 million in government funding and a private investment of £260 million), demonstrate how government and private investments can synergistically support projects that drive innovation and economic growth[357]. This collaborative model not only fuels technological advancements but also offers a platform for academia to engage in cutting-edge research while benefitting from industry resources.

Other partnerships could extend beyond financial aspects, encompassing joint research projects[358], shared publications, and intellectual exchanges at conferences or through informal networks[359]. They also offer practical engagement opportunities like internships and sabbaticals, allowing academics to gain industry experience without departing from their academic roles[360].

351 Wright B *et al.* 2014 Technology transfer: Industry-funded academic inventions boost innovation. *Nature* **507**, 297–299. https://doi.org/10.1038/507297a

352 *Ibid.*

353 Kunze L. 2019. Can we stop the academic AI brain drain? *KI-Künstliche Intelligenz*, *33*(1), 1-3. (https://doi.org/10.1007/s13218-019-00577-2)

354 Gofman M, Jin Z. 2022 Artificial Intelligence, Education, and Entrepreneurship. *Journal of Finance, Forthcoming*. (https://doi.org/10.1111/jofi.13302)

355 UK universities alarmed by poaching of top computer science brains. *Financial Times. 9 May 2018.* See https://www.ft.com/content/895caede-4fad-11e8-a7a9-37318e776bab (accessed 10 June 2023)

356 Life sciences companies supercharged with £277 million in government and private investment. Gov.UK See https://www.gov.uk/government/news/life-sciences-companies-supercharged-with-277-million-in-government-and-private-investment (accessed 26 February 2024)

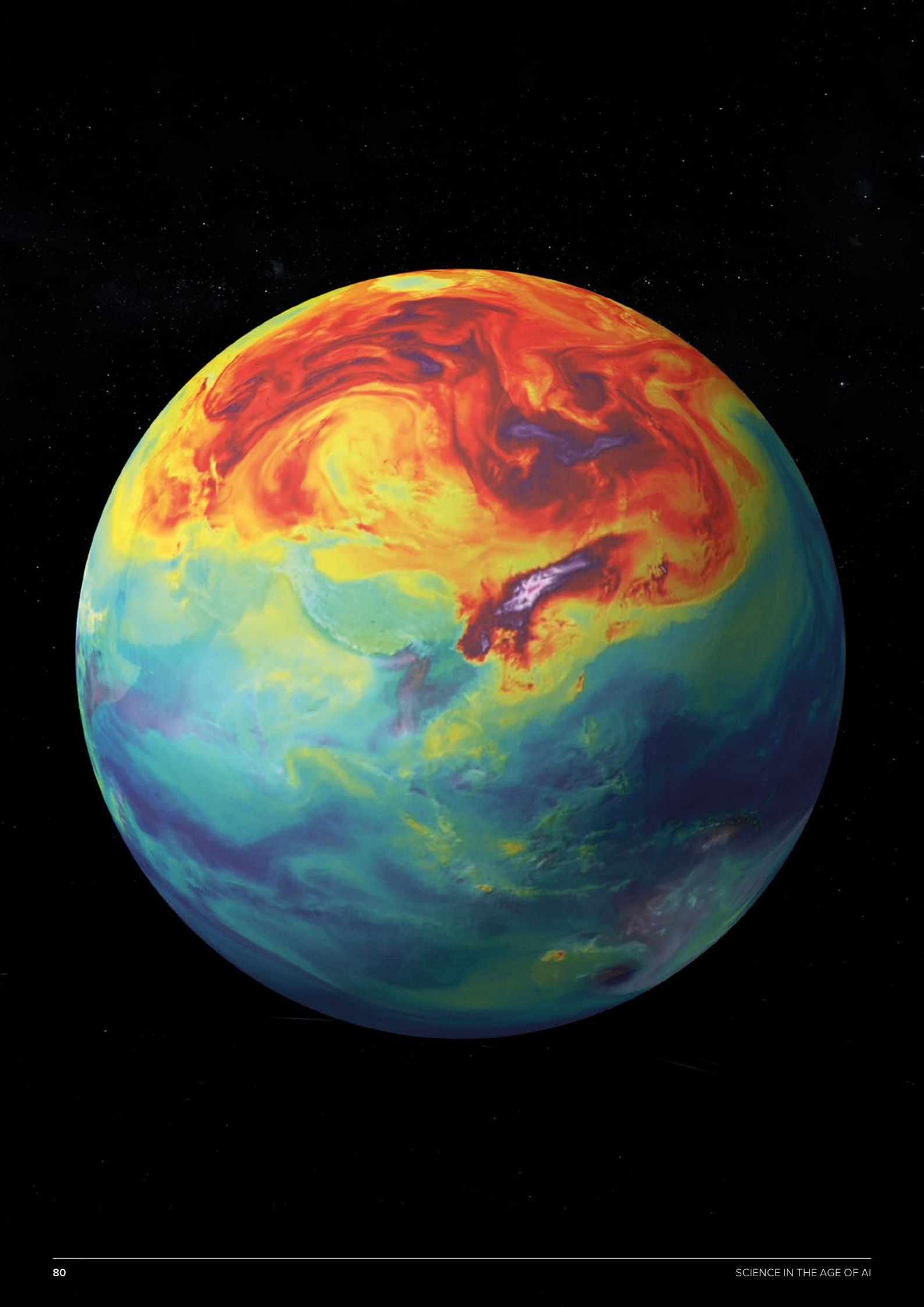
357 Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI. Gov.UK See https://www.gov.uk/government/news/initial-100-million-for-expert-taskforce-to-help-uk-build-and-adopt-next-generation-of-safe-ai (accessed 26 February 2023

358 Evans JA (2010) Industry induces academic science to know less about more. Am J Sociol 116(2):389–452

359 Perkmann M, Walsh K (2007) University–industry relationships and open innovation: towards a research agenda. Int J Manag Rev 9(4):259–280 (https://doi.org/10.1111/j.1468-2370.2007.00225.x)

360 Cohen WM, Nelson RR, Walsh JP. 2002 Links and impacts: the influence of public research on industrial R&D. Manag Sci 48(1):1–23.(https://doi.org/10.1287/mnsc.48.1.1.14273)

To optimise the benefits of cross-sector collaboration, universities and research institutions can also develop robust IP policies, fostering an environment where innovation is protected and can be shared effectively with industry partners. By creating structured pathways for collaboration, such as joint patenting efforts or licensing agreements, both sectors can contribute to advancing AI research while addressing challenges like resource disparities and data privacy.



Chapter five
Research ethics and AI safety

Left
Carbon dioxide emissions.
© iStock / janiecbros.

Research ethics and AI safety

"We might not want to make some of the datasets available because of the ease of misuse. And that seems sort of the opposite of what we try to strive for as scientists. But we may want to think more about safety and security now."

The Royal Society

As the use of AI expands across scientific disciplines, new ethical challenges are arising around the unintended or intended misuse of AI[364]. There is also a growing concern from the public regarding the fair and ethical use of their data[365] and the extent to which AI-based tools can propagate harmful biases, discrimination, and societal harms[366]. AI Safety risks also need to be considered as it has become easier to repurpose algorithms for malicious use[367].

Ahead of the Global AI Safety Summit hosted by the United Kingdom in 2023, the Royal Society and the UK's Department for Science, Innovation and Technology (DSIT) convened a horizon scanning workshop on AI safety. The following themes emerged as ethical challenges associated with the use of AI in scientific research:

1. **Data and algorithmic bias**
   AI systems can have biases embedded in them through training data and algorithmic design. When left unmitigated, algorithmic bias can lead to unfair outcomes and exacerbate inequalities[368]. The integration of AI in medicine has, for example, highlighted how algorithmic biases can lead to inaccurate medical diagnoses, inadequate treatment, and exacerbated healthcare disparities[369,370]. If data bias translates into the training data for AI models, there is a risk that models will not map well on to other communities[371] (See Box 6).

   Algorithmic harms can also manifest in the realm of funding and scholarly communication. AI tools are used to make initial screenings of grant and peer review processes less time intensive[372]. Among other applications, it can support reviewers in identifying false citations[373], boost the quality of papers[374] and reduce plagiarism[375].

364 Ghotbi N. 2024. Ethics of Artificial Intelligence in Academic Research and Education. In *Second Handbook of Academic Integrity* (pp. 1355-1366). (https://doi.org/10.1007/978-981-287-079-7_143-1)

365 Lomas N. 2023 UK court tosses class-action style health data misuse claim against Google Deepmind. Tech Crunch. 19 May 2023. See https://techcrunch.com/2023/05/19/uk-court-tosses-class-action-style-health-data-misuse-claim-against-google-deepmind (accessed 21 December 2023)

366 Brennan, J. 2023. AI assurance? Assessing and mitigating risks across the AI lifecycle. Ada Lovelace Institute. See https://www.adalovelaceinstitute.org/report/risks-ai-systems/ (accessed September 30 2023)

367 Urbina F, Lentzos F, Invernizzi C, Ekins S. 2022. Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence, 4(3), 189-191. (https://doi.org/10.1038/s42256-022-00465-9)

368 UK Parliament POST. 2024. Policy implications of artificial intelligence (AI). https://researchbriefings.files.parliament.uk/documents/POST-PN-0708/POST-PN-0708.pdf

369 Panch, T., Mattie, H. and Atun, R. 2019. Artificial intelligence and algorithmic bias: implications for health systems. *Journal of global health*, *9*(2).

370 Celi, L.A *et al*,. 2022. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health*, *1*(3), p.e0000022.

371 Royal Society and Department for Science, Innovation, and Technology workshop on horizon-scanning AI safety risks across scientific disciplines, 2023.

372 Checco A, Bracciale L, Loreti P, Pinfield S, Bianchi G. 2021 AI-assisted peer review. *Humanities and Social Sciences Communications*, **8**(1), pp.1-11. (https://doi.org/10.1057/s41599-020-00703-8)

373 The Royal Society roundtable on Large Language Models, July 2023

374 Checco A, Bracciale L, Loreti P, Pinfield S, Bianchi G. 2021 AI-assisted peer review. *Humanities and Social Sciences Communications*, **8**(1), pp.1-11. (https://doi.org/10.1057/s41599-020-00703-8)

375 Heaven D. 2022 AI peer reviewers unleashed to ease publishing grind. Nature. 22 November 2018. See https://www.nature.com/articles/d41586-018-07245-9 (accessed 27 March 2023)

While AI can contribute towards reducing "first-impression" bias by human reviewers, it can also reinforce pre-existing gender, language or institutional biases entrenched in training datasets[376] which can harm career progression opportunities for underrepresented scholars[377]. Initiatives like the GRAIL project are exploring ethical principles and best practices for using AI in research funding and evaluation[378].

BOX 6

Multilingual language models

The development of multilingual language models can contribute towards reducing biased outputs. For example, BLOOM (BigScience Language Open-science Open-access Multilingual) is currently the largest open research language model developed with the objective of reducing harmful and biased outputs by training it on a smaller selection of higher-quality, multilingual text sources. BLOOM was developed by 1000 researchers and trained on 46 different languages and 13 programming languages[379]. Training data is available, and the model has been developed under an ethical charter that centres values such as diversity, inclusivity, reproducibility; and aims to foster accessibility, multilingualism and interdisciplinarity[380].

2. Hallucinations and AI-generated disinformation

The growing use of general-purpose or foundation models in science (eg generative AI and LLMs) brings about unique considerations around ethics and safety. For example, while LLMs can be used to accelerate academic writing, they can also be used to intentionally generate scientific disinformation[381]. Increased public access to LLMs reduces barriers for malicious actors to generate convincing machine-created content that reduces the likelihood of human detection[382] (see Box 1 on the Royal Society's red teaming exercise on scientific disinformation).

The use of LLMs in a scientific project, can also increase exposure to 'hallucinations' – which refers to the generation of convincing and realistic outputs which do not correspond to real-world inputs. Even when there is no malicious intent, general pre-trained transformer (GPT) technologies can fabricate facts, data and citations when responding to a prompt. The rapid surge of machine-generated disinformation online increases the risk that the next generation of models trained on web-scraped data will degrade in performance and absorb distortions and inaccuracies found in fabricated text and data[383].

376 Checco A, Bracciale L, Loreti P, Pinfield S, Bianchi G. 2021 AI-assisted peer review. *Humanities and Social Sciences Communications*, **8**(1), pp.1-11. (https://doi.org/10.1057/s41599-020-00703-8)

377 Chawla, D.S. 2022, Should AI have a role in assessing research quality?. *Nature*. (DOI: 10.1038/d41586-022-03294-3)

378 Research on Research Institute. See https://researchonresearch.org/project/grail/ (accessed 5 January 2024)

379 Hugging Face. Documentation of BLOOM. See: https://huggingface.co/docs/transformers/model_doc/bloom (accessed 21 December 2023)

380 Hugging Face. BigScience Ethical Charter. See: https://bigscience.huggingface.co/blog/bigscience-ethical-charter (accessed 21 December 2023)

381 Wang H *et al*. 2023 Scientific discovery in the age of artificial intelligence. *Nature*, **620**. 47-60. (https://doi.org/10.1038/s41586-023-06221-2)

382 Bommasani *et al*. 2021. On the opportunities and risks of foundation models. See: https://crfm.stanford.edu/assets/report.pdf (accessed March 21 2024)

383 Pan Y, Pan L, Chen W, Nakov P, Kan M Y, Wang W Y. 2023. On the Risk of Misinformation Pollution with Large Language Models. *arXiv preprint (arXiv:2305.13661)*

For example, Meta's LLM for science, Galactica, was trained on 48 million scientific articles, websites, textbooks, and other inputs to help researchers summarise the literature, generate academic papers, write scientific code and annotate data (eg, molecules and proteins). However, the demo was paused after three days of use. One of the largest risks posed by Galactica was how confidently it produced false information and the lack of guidelines to identify it[384].

As with other forms of misinformation, hallucinations can erode public trust in science[385]. Methods for AI validation and disclosure, such as watermarking or content provenance technologies[386], are being explored to enable the detection of AI-generated content and mitigate potential harms caused by hallucinations[387],as well as, to ensure public trust in emerging AI systems[388].

3. Dual use of AI technologies developed for science

The dual use of AI systems refers to situations in which a system developed for a specific use is then appropriated or modified for a different use. Malicious use refers to applications in which the intent is to cause harm[389]. Among the most prominent and documented examples of malicious use of AI, is the development of chemical and biological weapons using AI systems that have beneficial applications for scientific research.

In 2020, the company Collaborations Pharmaceuticals, a biopharma company that builds ML models to assist drug discovery and the treatment of rare diseases, published results on what they have called a 'teachable moment' regarding the use of AI-powered drug discovery methods. Following an invitation from the Swiss Federal Institute for NBC (nuclear, biological, and chemical) protection, the company trained an AI-powered molecule generator used for drug discovery to generate toxic molecules within a specified threshold of toxicity[390]. Drawing from a public database, and in less than 6 hours, the model had generated 40,000 molecules. Many of these molecules were similar or more toxic than the nerve agent VX, a banned and highly toxic lethal chemical weapon.

While the theoretical generation of toxic molecules does not imply their production is viable or feasible, the experiment shows how AI can speed up the process of creating hazardous substances, including lethal bioweapons[391]. The company has

384 MIT Review. Why Meta's latest large language model survived only three days online 2022. See: https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ (accessed September 30 2023)

385 Bontridder, N. and Poullet, Y., 2021. The role of artificial intelligence in disinformation. *Data & Policy*, *3*, p.e32. (doi:10.1017/dap.2021.20)

386 The Royal Society. Generative AI, content provenance and a public service internet. See: https://royalsociety.org/news-resources/publications/2023/digital-content-provenance-bbc/

387 Watermarking refers to techniques that can embed identification information into the original data, model, or content to indicate provenance or ownership.

388 Partnership on AI. PAI's Responsible Practices for Synthetic Media. See: https://syntheticmedia.partnershiponai.org/#read_the_framework (accessed 21 December 2023)

389 Ueno, H. 2023. Artificial Intelligence as Dual-Use Technology. In *Fusion of Machine Learning Paradigms: Theory and Applications* (pp. 7-32). Cham: Springer International Publishing. (https://doi.org/10.1007/978-3-031-22371-6_2)

390 Urbina F, Lentzos F, Invernizzi C, Ekins S. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189-191. (https://doi.org/10.1038/s42256-022-00465-9)

391 Sohn R. 2022.AI Drug Discovery Systems Might Be Repurposed to Make Chemical Weapons, Researchers Warn. Scientific American [Internet]. 21 April 2022. See https://www.scientificamerican.com/article/ai-drug-discovery-systems-might-be-repurposed-to-make-chemical-weapons-researchers-warn/ (accessed 21 December 2023)

called for several actions to address security risks, as well as to monitor, address and limit malicious applications of AI models used in science[392].

4. Data poisoning and adversarial machine learning attacks

Measures to enhance AI safety in science need to consider the development of robust models that can withstand adversarial data-based attacks[393]. Training AI models on vast, and poorly curated datasets, creates vulnerabilities to instances of 'data poisoning', 'false data injections' or 'one-pixel' attacks[394].

These tactics involve inserting noisy, incorrect, or manipulated data to deceive machine learning systems while remaining imperceptible or hard to detect for humans[395,396,397]. 'Poisoned' or manipulated datasets are one of the most common and documented attacks to the reliability of AI systems[398].

5. Environmental costs of using AI systems

The collection, analysis, storage and sharing of data required for AI-based systems has a significant environmental impact[399]. For example, storing a terabyte of data is estimated to consume 10kg of carbon dioxide annually[400], while training a ChatGPT-style LLM can create 550 tonnes of carbon dioxide emissions[401]. It is estimated that the global greenhouse gas emissions of data centres are the same as the emissions of US commercial aviation, and as datasets and models get larger, this is likely to increase.

To mitigate the negative impacts of climate change, these systems will need to meet the principle of energy proportionality[402] and environmentally sustainable computational science (ESCS) best practices. Other developments to improve the environmental sustainability of AI-based tools include:

- Integration of green computing techniques into research methods[403]

392 Urbina F, Lentzos F, Invernizzi C, Ekins S. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189-191. (https://doi.org/10.1038/s42256-022-00465-9)

393 Collins K *et al*. 2023. 'Human Uncertainty in Concept-Based AI Systems.' Paper presented at the *Sixth AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES 2023)*, August 8-10, 2023. Montréal, QC, Canada.

394 Su J, Vargas D V, Sakurai K. 2019 One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), pp.828-841. (DOI: 10.1109/TEVC.2019.2890858)

395 Verde L, Marulli F, Marrone S. 2021 Exploring the impact of data poisoning attacks on machine learning model reliability. *Procedia Computer Science*, *192*, pp.2624-2632. (https://doi.org/10.1016/j.procs.2021.09.032)

396 Xu Y at al. 2021 Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, *2*(4). (https://doi.org/10.1016/j.xinn.2021.100179)

397 Liu Y, Ning P, Reiter M K. 2011 False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)*, *14*(1), pp.1-33. (https://doi.org/10.1145/1952982.1952995)

398 Verde L, Marulli F, Marrone S. 2021 Exploring the impact of data poisoning attacks on machine learning model reliability. *Procedia Computer Science*, *192*, pp.2624-2632. (https://doi.org/10.1016/j.procs.2021.09.032)

399 Henderson P, Hu J, Romof J, Brunskill E, Jurafsky D, Pineau J (2020) Towards the systematic reporting of the energy and carbon footprints of machine learning. J Mach Learn Res 21(248):1–43

400 Lannelongue L *et al.* 2023 GREENER principles for environmentally sustainable computational science. *Nat Comput Sci* 3, 514–521. (doi.org/10.1038/s43588-023-00461-y)

401 Patterson D *et al*. 2021 Carbon emissions and large neural network training. *arXiv preprint*. (doi.org/10.48550/arXiv.2104.10350)

402 The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/ (accessed 21 December 2023).

403 Chithra, J, Vijay, A, & Vieira, D 2014. A study of green computing techniques. Int. J. Comput. Sci. Inf. Technol.

- Certification standards for sustainable lab practices. Expanding 'Green Lab' grassroots networks in conjunction with institutional policies can reshape researcher norms[404].
- Funders can mandate carbon reporting and support upgrades like energy-efficient hardware in AI-based research projects[405].
- Governmental mandates and regulatory pressure – the EU's carbon reporting rules now cover 50,000+ companies[406].
- Using AI to optimise and minimise the environmental impact of research methods. For example, immersive technologies provide virtual experiences, minimising on-site damage and visualisation which can bolster preparation efficiency[407,408].

6. Human cost of training AI systems

The development and use of AI tools relies on a critical but often invisible human infrastructure. Even though human labour is essential for AI deployment, in some cases it remains underappreciated under the guise of 'automation'. Technology critic, Astra Taylor argues that the discourse of automation can be used to marginalise certain contributors to the scientific process (eg women or ghost workers) and justify cost-cutting measures without addressing issues of equity and fairness[409,410].

The impact on labour can span from shifts in the labour market[411] to the exploitation of data workers that power large AI systems[412]. Interrogating the organisation of labour can contribute towards generating accountability to develop responsible AI supply chains[413].

Addressing AI ethics in scientific research

There are opportunities for the scientific community (from scientists to system developers and funders)[414] to proactively consider strategies to monitor, anticipate and respond to unforeseen harms caused by the use of AI systems[415].

404 Green Your Lab Network. See: https://network.greenyourlab.org/ (accessed March 21 2024)

405 The Royal Society roundtable on AI and climate science, June 2023.

406 More than 50,000 companies to report climate impact in EU after pushback fails. Financial Times. 18 October 2024. See: https://www.ft.com/content/a3216188-8e50-4a62-a8d9-e89172b3ddc7 (accessed March 21 2024)

407 The Royal Society. 2023 Science in the metaverse: policy implications of immersive technologies. See https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/ (accessed 21 December 2023).

408 The Royal Society interviews with scientists and researchers. 2022 - 2023

409 Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

410 Taylor A. 2018 The Automation Charade. LogicMag. 1 August 2018. See: https://logicmag.io/05-the-automation-charade/ (accessed February 28 2024)

411 World Economic Forum, These are the jobs most likely to be lost – and created – because of AI. See: https://www.weforum.org/agenda/2023/05/jobs-lost-created-ai-gpt/ (accessed February 24 2024)

412 Barret, M. The dark side of AI: algorithmic bias and global inequality. See: https://www.jbs.cam.ac.uk/2023/the-dark-side-of-ai-algorithmic-bias-and-global-inequality/ (accessed December 10 2023)

413 Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

414 Wang H *et al*. 2023 Scientific discovery in the age of artificial intelligence. *Nature*, **620**. 47-60. (https://doi.org/10.1038/s41586-023-06221-2)

415 Kazim E, Koshiyama A S. 2021 A high-level overview of AI ethics. *Patterns*, *2*(9). (https://doi.org/10.1016/j.patter.2021.100314)

Drawing from interviews and roundtable discussions, the following measures were suggested to ensure the ethical application of AI across sectors:

- Domain-specific taxonomy for harms: Establish audits, impact assessments or evaluation frameworks to understand socio-technical harms stemming from different fields. Examples include the taxonomy published by DeepMind listing different types of human-computer interaction, environmental and socioeconomic harms[416]. Another example is the multi-stakeholder framework developed to evaluate the Social Impact of Generative AI in Systems and Society. It accounts for harms related to representation, cultural values and sensitive content, performance, privacy and data protection, financial costs, environmental costs, and labour costs[417].
- Ethical guidelines and reviews: Ethical guidelines and codes of conduct can guide design of AI models used for science, prevent misuse and establish best practices. Examples include Hague Ethical Guidelines[418], which promote a code of conduct to guard against the misuse of chemistry research, or UNESCO's guidelines for Ethical Artificial Intelligence, the first-ever global standard on AI ethics aimed at maximising the benefits and minimising the downside risks of the use of AI for scientific discoveries[419].

Further domain-specific guidance is needed to ensure scientists across domains and sectors can make informed decisions when integrating AI into their work.

- Communication and knowledge sharing: Drawing from the United Nations Office of Disarmament Affairs[420], US-based private sector companies (OpenAI, Anthropic, Microsoft, Hugging Face), and civil society have put forward a proposal to improve trust through confidence-building measures, such as communication and coordination, observation and verification, cooperation and integration, and transparency[421].
- Sanctions and restrictions: Explore the regulation of specific software and applications in industry and academia, and the viability of limiting access to tools and models with high potential for misuse[422].
- Public engagement: Explore new governance approaches to engage affected publics in the co-construction of constraints and guardrails. A strategy to communicate risk to the public also needs to be considered, while preventing a general loss of trust in science.

416 Weidinger L *et al*. 2022 Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214-229). (https://doi.org/10.1145/3531146.3533088)

417 Solaiman, I *et al*. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint (arXiv:2306.05949*)

418 The Organisation for the Prohibition of Chemical Weapons (OPCW). The Hague Ethical Guidelines. See: https://www.opcw.org/hague-ethical-guidelines (accessed 28 February 2024)

419 UNESCO Recommendation on ethics of artificial intelligence. 2022. See: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence (accessed 6 February 2024)

420 United Nations, Office of Disarmament Affairs. Confidence Building Measures. See: https://disarmament.unoda.org/biological-weapons/confidence-building-measures/ (accessed 21 December 2023)

421 Shoker S *et al*. 2023 Confidence-building measures for artificial intelligence: Workshop proceedings. *arXiv preprint (arXiv:2308.00862*)

422 Urbina F, Lentzos F, Invernizzi C, Ekins S. 2023 Preventing AI From Creating Biochemical Threats. *Journal of Chemical Information and Modeling*, *63*(3), 691-694 (https://doi.org/10.1021/acs.jcim.2c01616)

CASE STUDY 3

AI and climate science

As AI and ML capabilities become further integrated into climate science research and applications, they are expanding the capacity of scientists and policy makers to mitigate the climate crisis[423].

Opportunities for AI in climate science research

Realising the potential of climate science data can be difficult due to the heterogeneous nature of environmental data. This presents a challenge in linking data together due to misaligned data across spatiotemporal resolutions; varying privacy and security levels (particularly in relation to satellite imaging); and a lack of regulation[424].

DL techniques can interpolate measurement gaps for intricate pattern recognition in fields like space weather forecasting, a technique used in NASA's Centre for Climate Simulation[425, 426].

If done successfully, this fusion of datasets can improve the accuracy of models and estimates, contributing to long-term weather forecasting, which supports disaster preparedness and resource management for extreme events[427].

Other AI techniques have demonstrated effectiveness in forecasting global mean temperature changes[428], predicting climatic phenomena like El Niño[429], cloud systems[430], and regional weather patterns, such as rainfall in specific areas[431]. For instance, a 2023 *Nature* paper showed an AI model had predicted weather better than the world's most advanced forecasting system[432], soon after, DeepMind's ML approach surpassed even that benchmark[433]. The Royal Society's 2020 report, *Digital technology and the planet: Harnessing computing for net zero*, also outlines the role AI can play in achieving global net zero ambitions[434].

423 Huntingford C, Jeffers E S, Bonsall M B, Christensen H M, Lees T, Yang H. 2019 Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters, 14(12), 124007. (DOI 10.1088/1748-9326/ab4e55)

424 National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/27520.

425 Kadow, C, Hall, DM, Ulbrich, U, 2020. Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, **13**, pp.408-413. (https://doi.org/10.1038/s41561-020-0582-5)

426 NASA Centre for Climate Simulation. See https://www.nccs.nasa.gov/news-events/nccs-highlights/acceleratingScience. (Accessed 21 December 2023)

427 Buizza, C *at al* 2022. Data learning: Integrating data assimilation and machine learning. *Journal of Computational Science*, *58*, p.101525. (https://doi.org/10.1016/j.jocs.2021.101525)

428 Ise, T, Oba, Y. 2019. Forecasting climatic trends using neural networks: an experimental study using global historical data. *Frontiers in Robotics and AI*, **32**. (https://doi.org/10.3389/frobt.2019.00032)

429 Ham, YG, Kim, JH, Luo, JJ. 2019. Deep learning for multi-year ENSO forecasts. *Nature*, **573**, 568-572. (https://doi.org/10.1038/s41586-019-1559-7)

430 Rasp, S, Pritchard, MS., & Gentine, P. 2018. Deep learning to represent sub grid processes in climate models. *Proceedings of the National Academy of Sciences*, **115**, 9684-9689. (https://doi.org/10.1073/pnas.181028611)

431 Zheng, G, Li, X, Zhang, RH, Liu, B 2020. Purely satellite data–driven deep learning forecast of complicated tropical instability waves. *Science advances*, **6**, eaba1482. (DOI: 10.1126/sciadv.aba1482)

432 Bi, K, Xie, L, Zhang, H, Chen, X, Gu, X, Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533-538. (https://doi.org/10.1038/s41586-023-06185-3)

433 Wong, C. 2023. DeepMind AI accurately forecasts weather-on a desktop computer. *Nature*. 14 November 2023 (https://doi.org/10.1038/d41586-023-03552-y)

434 The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/ (accessed 21 December 2023).

AI for climate preparedness and decision-making

Adopting a systems approach[435] to AI in climate science could enable more effective climate decision-making[436]. AI-driven climate models not only simulate complex systems, but reinforced learning can also contribute to sustainable policymaking through the systematic evaluation of climate actions, trade-offs, and risks[437]. An example of this includes digital twins, a virtual representation of a physical asset which can be used to understand, predict, and optimise the performance of this asset[438].

The European Space Agency's Destination Earth is developing a digital twin that can generate rich data flows to enable a 'control loop' for the planet's emissions[439]. This can facilitate monitoring of the natural and human activity contributing to climate change, allow experts to anticipate and plan for extreme events and adapt policies to climate related challenges[440]. Real-world applications, like Tuvalu's digital twin of the island nation, to safeguard its existence against sea level rise demonstrates the promise and the vital need for this technological development[441].

435 The Royal Academy of Engineering 2020 Net Zero: A systems perspective on the climate challenge. See raeng.org.uk/publications/reports/net-zero-a-systems-perspective-on-the-climate-chal (accessed 14 October 2020)

436 The Royal Society. 2021 Computing for net zero: how digital technology can create a 'control loop for the protection of the planet'. See https://royalsociety.org/-/media/policy/projects/climate-change-science-solutions/climate-science-solutions-computing.pdf (accessed 21 December 2023)

437 Abrell J, Kosch M, Rausch S (2019) How effective was the UK carbon tax?—A machine learning approach to policy evaluation. SSRN Scholarly Paper ID 3372388. Social Science Research Network, Rochester. 10.2139/ssrn.3372388

438 The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/ (accessed 21 December 2023)

439 The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/ (accessed 21 December 2023).

440 The European Space Agency. Destination Earth. See https://www.esa.int/Applications/Observing_the_Earth/Destination_Earth. (accessed 21 December 2023)

441 Accenture. Case Study: Tuvalu. See https://www.accenture.com/us-en/case-studies/technology/tuvalu. (accessed 21 December 2023)

Research ethics challenges in climate science

AI-based climate science holds great potential for solving global climate change challenges. However, this potential comes with challenges to research ethics. Identifying the risks and minimising pitfalls remains vital to maximising positive impact.

1. **Environmental costs of AI**
   Running complex simulations can have high carbon footprints from the immense computational power required, potentially offsetting intended environmental benefits of climate science research if sustainability practices are not integrated[442].

2. **Data bias and inaccuracies**
   Bias and inaccuracies in training data presents another key challenge. For instance, models disproportionately trained on Western populations risk overlooking issues in the Global South and entrenching unfairness for disadvantaged regions. Ongoing participation, auditing, and updating of systems with localised data is critical for equity[443].

3. **Global funding and grants**
   There is a significant global disparity in funding and grant distribution. An analysis into energy and climate research funding between 1990 and 2020 global disparity in funding distribution with Western countries (European Commission, UK, and US) receiving most of the funding[444]. Notably, the paper found that no research institution from Africa ranked among the top 10 funded institutions. Increased investment in AI initiatives for underrepresented regions is needed to support capacity-building and fostering of climate scientists in the Global South[445].

4. **Sensitive data sharing**
   Environmental data can contain sensitive information, include private or personal details that can be linked to specific, nonconsenting individuals or communities[446]. High-resolution spatial data, and digital traces pose privacy risks, which are magnified when researchers lack cultural understanding and sensitivity towards different communities. For example, while well intentioned, the digitisation of land records to increase researcher access to data can result in private actors like landowners with more financial resources to capitalise on this new data[447].

442 Henderson P, Hu J, Romof J, Brunskill E, Jurafsky D, Pineau J 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. J Mach Learn Res **21**:1–43

443 Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See https://royalsociety.org/current-topics/ai-data/ (accessed 7 May 2024)

444 AbdulRafiu A, Sovacool B K, Daniels C. 2022 The dynamics of global public research funding on climate change, energy, transport, and industrial decarbonisation. *Renewable and Sustainable Energy Reviews*, **162**, 112420. (https://doi.org/10.1016/j.rser.2022.112420)

445 Grantham Research Institute on Climate Change and the Environment. What opportunities and risks does AI present for climate action? See: https://www.lse.ac.uk/granthaminstitute/explainers/what-opportunities-and-risks-does-ai-present-for-climate-action/

446 Zipper S C *et al*. 2019 Balancing open science and data privacy in the water sciences. *Water Resources Research*, **55**, 5202-5211.(https://doi.org/10.1029/2019WR025080)

447 Donovan K P. 2012 Seeing like a slum: Towards open, deliberative development. *Georgetown Journal of International Affairs*.

Environmentally sensitive data can also adversely impact the environment[448]. For example, sharing biodiversity data, such as nesting locations of rare birds, can lead to bad actors harming those environments[449].

Strategies for ethical AI-based research practices in climate science

1. **Pursuing energy proportionality**
   Develop strategies to ensure that technologies developed in pursuit of net zero deliver environmental benefits that outweigh their emissions[450]. Interdisciplinary research on carbon accounting and impact assessment tools like the Green Algorithms Project[451] can contribute towards evaluating and mitigating the environmental impact of computational processes used in climate science.

2. **Improving global researcher access to data**
   The disparity in researcher access to data raises concerns about the equitable development and application of AI[452]. This could hinder the development of effective climate solutions tailored to the unique challenges of specific communities. Networks such as the Pacific Community's Statistics for Development Division can promote equitable access to data across diverse contexts, fostering collaboration and knowledge sharing[453]. Similarly, the establishment of trusted data institutions can contribute towards enhancing data sharing and usage to address emergencies and crises[454] [455].

3. **Contextualising data governance**
   Universal approaches to open data do not always engage with minority groups' rights and interests. Existing data sharing principles like FAIR (findable, accessible, interoperable, reusable)[456] can be complemented by people and purpose-oriented governance principles like the CARE Principles for Indigenous Data Governance (collective benefit, authority to control, responsibility, ethics) that considers a broader approach to sensitive data[457].

448 NBN. Sensitive Data. See https://nbn.org.uk/sensitive-data/ (accessed 6 March 2024)

449 *Ibid.*

450 The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/ (accessed 21 December 2023).

451 Green Algorithms Project. See https://www.green-algorithms.org/ (accessed 21 December 2023)

452 National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/27520.

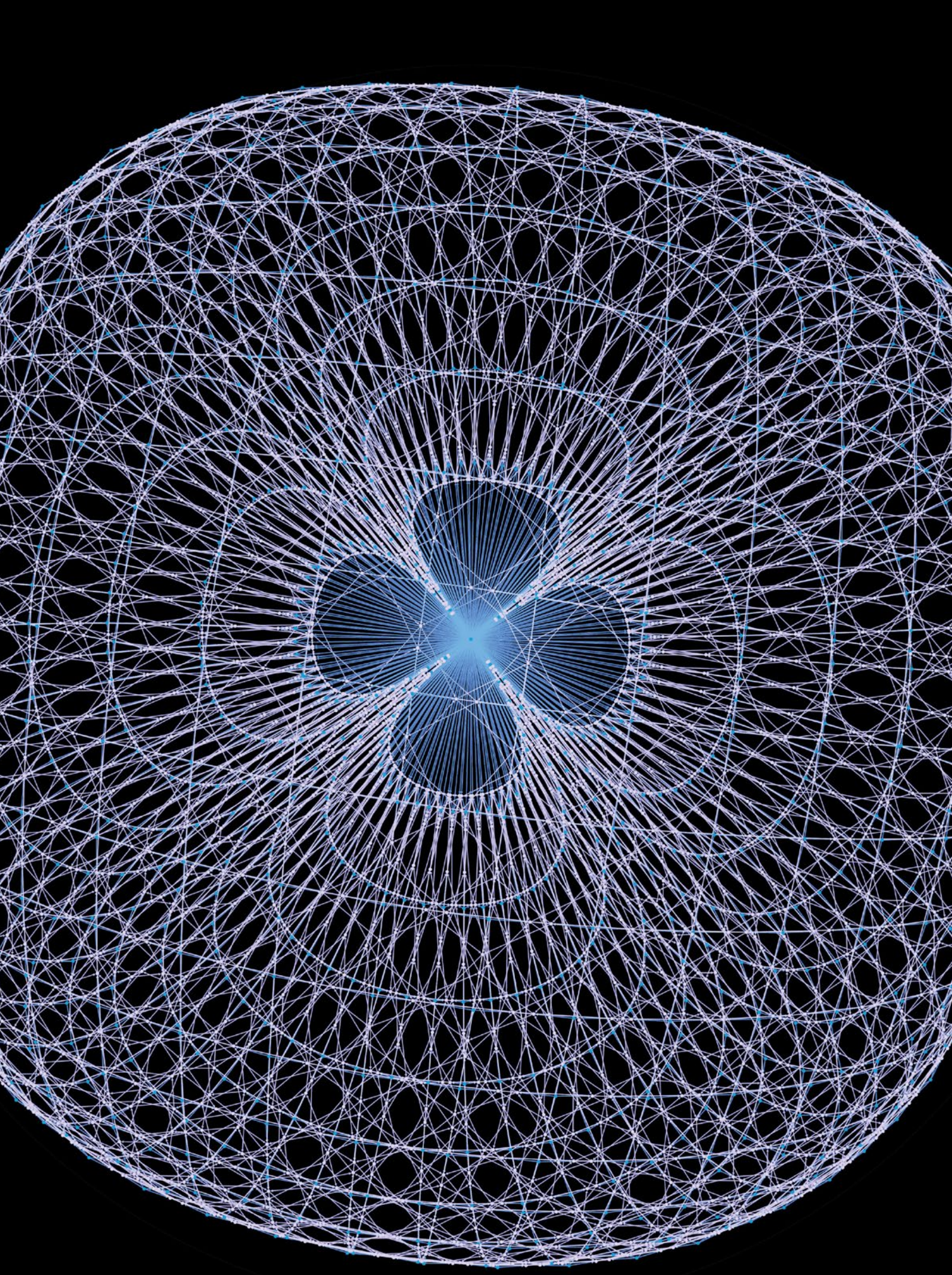
453 Pacific Community. Statistics for Development Division. See https://sdd.spc.int/ (accessed 6 March 2024)

454 The Royal Society. 2023 Creating resilient and trusted data systems. See https://royalsociety.org/topics-policy/projects/data-for-emergencies/ (accessed 21 December 2023).

455 Global Partnership on Artificial Intelligence. 2023 Designing Trustworthy Data Intuitions - Scanning the Local Data Ecosystem in Climate-Induced Migration in Lake Chad Basin - Pilot Study in Cameroon. See: https://gpai.ai/projects/data-governance/ (accessed 6 March 2024)

456 Wilkinson M. D., *et al.* 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9. (doi: 10.1038/sdata.2016.18.)

457 Global Indigenous Data Alliance. Care Principles for Indigenous Data Governance. See https://www.gida-global.org/care (accessed 21 December 2023)



Conclusion

Left
Graphcore Stereo Image Matching Benchmark, October 2015.

Conclusion

As explored throughout the report, the applications of AI in scientific research are bringing a new age of possibilities and challenges. The transformative potential of AI, fuelled by big data and advanced techniques, offers substantial opportunities across domains. From mapping deforestation to aiding drug discovery and predicting rare diseases, the applications are vast and promising. Through the case studies on climate science, material science, and rare disease diagnosis, this report envisions a future in which AI can be a powerful tool for scientific researchers.

However, these opportunities bring about a series of challenges related to reproducibility, interdisciplinary collaboration, and ethics. Finding a balance in which scientists can harness the benefits of automation and the accelerated pace of discovery while ensuring research integrity and responsible use of AI will be essential. Following the Royal Society's commitment to ensuring science – and this case AI – is applied for the benefit of humanity, the report calls for collective efforts in addressing these challenges.

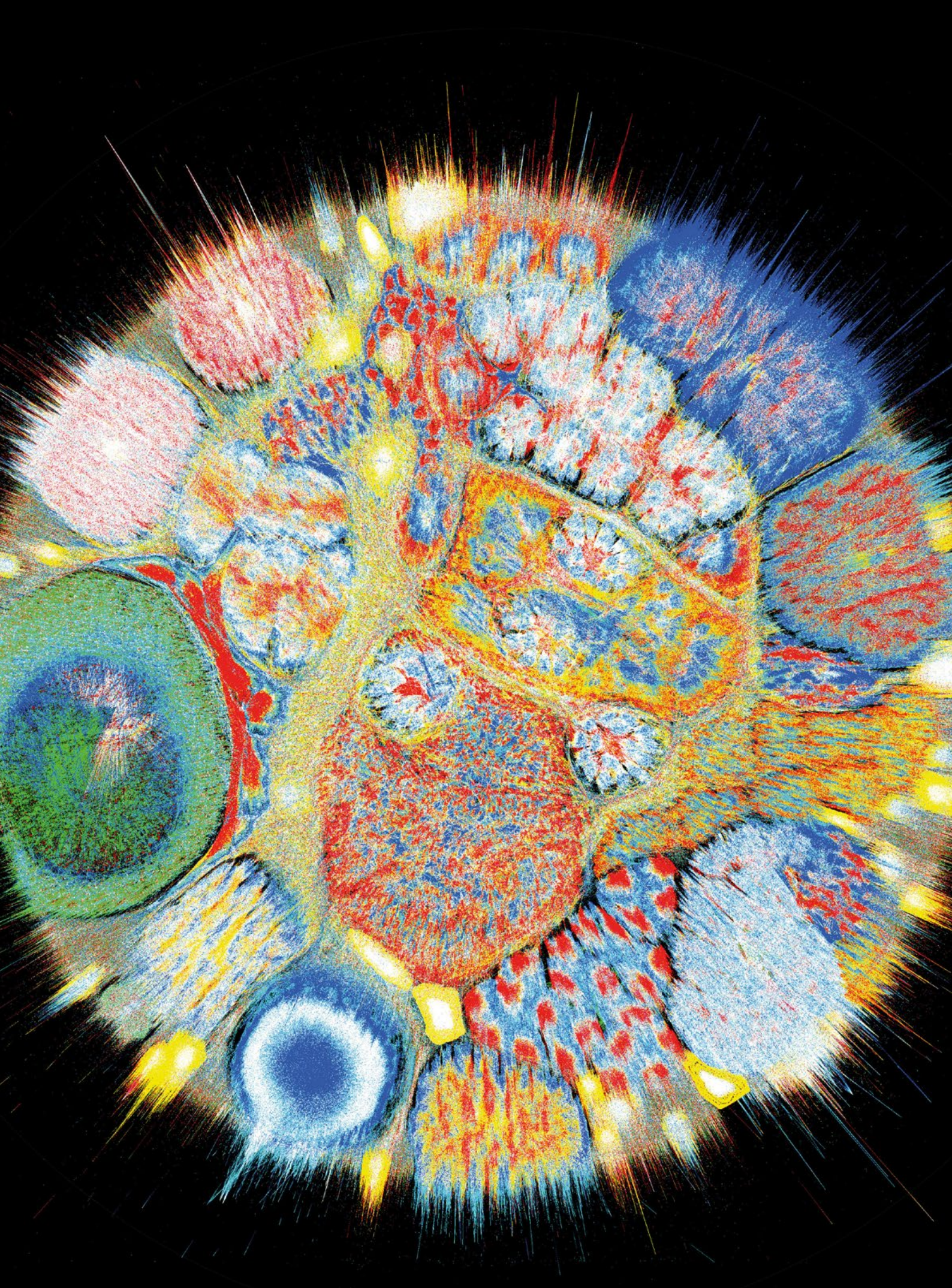
Moving forward, and according to the findings of this report, three areas of action require attention from scientific communities and relevant policy makers.

The first is to address issues of access and capability to use AI in science. Access to computing resources, high quality datasets, AI tools and relevant expertise is critical to achieve scientific breakthroughs. At the time of publication, access to essential infrastructures remained unequally distributed. This, coupled with a growing influence of the private sector as highlighted in Chapter 4 can have implications on the future of university-based AI research. Another challenge in this area is knowledge siloes between AI experts and scientific domain experts (Chapter 3). To ensure equitable distribution of AI across research communities, actions need to go beyond facilitating access, and focus on enhancing capabilities to collaborate, co-design and use AI across different scientific fields and research environments.

Second, open science principles and practices offer a clear pathway to improve transparency, reproducibility, and public scrutiny – all of which have proven challenging in AI-based scientific projects. As stressed in Chapter 2, the stakes of not addressing these issues are high, posing risks not just to science but also to society if the deployment of unreliable or erroneous AI-based outputs leads to harms. Further work is needed to understand the interactions between open science and AI for science and how to best minimise safety and security risks stemming from the open release of models and data.

Third, as AI's role expands in science, ethical and safety considerations need to be centred in its design and implementation (Chapter 5). The growing reliance on large datasets, prompts questions about the potential misuse of sensitive information and biases that could perpetuate inequalities or lead to incorrect conclusions. The autonomous nature of AI systems also introduces safety risks, especially in fields like healthcare or environmental monitoring, where errors could have severe consequences; or in fields such as chemistry and biology, where datasets and models can be repurposed with malicious intent. Addressing these challenges requires interdisciplinary collaboration and building scientists' capacity to anticipate risk and provide oversight that minimises potential harms.

Looking ahead, further exploration by the scientific community and policymakers is needed to understand the implications of AI on the future of science. Questions about how universities can adapt training and skill requirements, how funders can continue to support non-AI scientific work and how to optimise AI for environmental sustainability are key to understand the impact of this on technology in science, society, and on the planet.



Appendices

Left
Microsoft Research ResNet-18
Training, April 2017.

APPENDIX 1

List of figures and boxes

APPENDIX 2

Further details on methodology

Summary of research activities

- Three commissioned research projects including a historical review on the role of disruptive technologies in transforming science, a taxonomy of the use of artificial intelligence in science, technology, engineering and medicine, and a patent landscape review of artificial intelligence and related inventions.
- 30+ semi-structured interviews
- Four roundtables on the topics of reproducibility, interdisciplinarity, climate science research and the impact of large language models (LLMs) in science.
- Horizon scanning exercise on AI risks for science co-organised with the Department of Science, Innovation and Technology (DSIT)
- International US-UK Scientific Forum on Researchers Access to Data, co-hosted by the Royal Society and the National Academy of Science

Commissioned evidence-gathering and reviews

- Penn J, 2024. *Historical review on the role of disruptive technologies in transforming science and society*.
- Berman B, Chubb J, and Williams K, 2024. *The use of artificial intelligence in science, technology, engineering, and medicine*.
- IP Pragmatics, 2024. *Artificial intelligence related inventions*.

Event and research activities

The Royal Society would like to thank all those who contributed to the development of this project, in particular through participation in the following events.

30+ interviews, August 2022 – June 2023

Royal Society staff interviewed scientists and researchers across scientific disciplines on emerging themes and technologies in their fields.

Name	Organisation
Aishik Ghosh	University of California Berkley
William Hersh	Oregon Health & Science University
Bruce Weir FRS	University of Washington
Charlotte Deane	University of Oxford
Charlotte Williams FRS	University of Oxford
Conan Donelly	International Niemann-Pick Disease Registry
David Gao	Nanolayers
Don Canfield ForMemRS	University of Southern Denmark
Doreen Cantrell FRS	University of Dundee
Eileen Furlong FRS	European Molecular Biology Laboratory (EMBL)
Emma Karoune	The Alan Turing Institute
Fran Platt FRS	University of Oxford
Frank Close FRS	University of Oxford
Jamie Rossjohn FRS	Monash University
Jane Francis FRS	British Antarctic Survey
Maria Perez-Ortiz	University College London
Mathieu Denis	International Science Council
Odd Erik Gunderson	Norwegian University of Science and Technology; Aneo
Paul Bates FRS	University of Bristol
Peter Dayan FRS	Max Planck Institute
Peter Hore FRS	University of Oxford
Philip Quinlan	University of Nottingham
Richard Benton FRS	University of Lausanne
Richard Horne FRS	British Antarctic Survey
Robin Franklin FRS	Altos Labs Cambridge Institute
Sean Ekins	Collaborations Pharmaceuticals
Simon Boulton FRS	Francis Crick Institute
Stephen Benkovic ForMemRS	Pennsylvania State University
Stephen Smartt FRS	University of Oxford
Tzung-Chien Hsieh	Universitat Bonn

***Roundtable on immersive technologies in scientific research*, June 2022**

The Royal Society hosted a roundtable at the University of Exeter, as part of the *Creating Connections* events series that convened academics and industry professionals from the Southwest of England to discuss the policy priorities for the use of immersive technologies in scientific research. The roundtable was chaired by Professor Samuel Vine, Professor of Psychology at the University of Exeter. The key topics discussed were the challenges faced by industry and academic researchers working with immersive technologies including training, sustainability, and access to markets. (See Royal Society website for more information)[458].

Name
Gavin Buckingham
Toby de Burgh
Neill Campbell
Kirsten Carter
Danae Stanton Fraser
Stephen Hobbiger
Nathan Mayne
Verity McIntosh
Stephanie Owens
Samuel Vine
Brain Waterfield
Carina Westling

458 The Royal Society. 2023 Science in the metaverse: policy implications of immersive technology. See https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/

***Roundtable on reproducibility*, April 2023**

The Royal Society's roundtable on the challenges of reproducibility in AI-based scientific research provided insights from Professor Sabina Leonelli and Joelle Pineau, and multiple reproducibility,

Name	Organisation
Dorothy Bishop FRS	University of Oxford
Joaquin Vanschoren	Eindhoven University of Technology; OpenML
Joelle Pineau	McGill University; Meta AI
Malvika Sharan	Alan Turing Institute; Open Life Science
Mark Kelson	University of Exeter
Odd Erik Gunderson	Norwegian University of Science and Technology; Aneo
Ralitsa Madsen	University of Dundee; UK Committee on Research Integrity
Rebecca Kirk	PLOS
Sabina Leonelli	University of Exeter
Sayash Kapoor	Princeton University
Susanna-Assunta Sanson	University of Oxford
Victoria Moody	JISC

***Roundtable on AI and climate science*, June 2023**

The Royal Society convened a roundtable for climate and data scientists to explore the role of AI in climate science research, share insights, challenges, and innovative ideas for the future of this field. Dame Professor Jane Francis FRS provided insights from role as director of the British Antarctic Survey.

Name	Organisation
Alistair Nolan	Organisation for Economic Co-operation and Development (OECD)
Anil Madhavapeddy	University of Cambridge
Anna Hogg	University of Leeds
Anna-Louise Ellis	Met Office
Dave Topping	University of Manchester
Emily Shuckburgh	University of Cambridge
Jane Francis FRS	British Antarctic Survey
Joycelyn Longdon	University of Cambridge; Climate in Colour
Konstantin Klemmer	Microsoft Research
Philip Stier	University of Oxford
Richard Turner	University of Cambridge
Scott Hosking	British Antarctic Survey; The Alan Turing Institute
Tommaso Venturini	University of Geneva; CNRS
Shakir Mohamed	Google DeepMind
Suman Ravuri	Google DeepMind
Richard Horne	British Antarctic Survey
Timothy Palmer	University of Oxford
Marian Scott	University of Glasgow

***Roundtable on interdisciplinarity*, July 2023**

The Royal Society convened a roundtable on the role of interdisciplinarity in AI-driven scientific research. The roundtable provided a comprehensive exploration of interdisciplinarity's pivotal role in navigating the transformative landscape of AI-driven scientific research, featuring insights from Professor Alison Noble FRS, prominent experts and organisations across academia and the private sector

Name	Organisation
Alison Noble FRS	University of Oxford
Alistair Nolan	Organisation for Economic Co-operation and Development (OECD)
Ankit Agrawal	Northwestern University
Bradley Love	University College London
Cecilia Mascolo	University of Cambridge
Claude Chelala	Queen Mary University of London
Daniele Quercia	King's College London; Nokia Bell Lab Cambridge
Gareth Conduit	University of Cambridge
Georgios Leontidis	The University of Aberdeen
Hujun Yin	University of Manchester
James Dracott	UKRI
Matthias Rillig	Freie Universität Berlin
Michael Castelle	University of Warwick
Mirco Musolesi	University College London
Raffaella Mulas	Vrije Universiteit Amsterdam
Reuben Shipway	University of Plymouth
Seth Baum	Global Catastrophic Risk Institute
Tommaso Venturini	University of Geneva; CNRS
Verena Reiser	Google DeepMind
Victoria Henickx	KU Leuven

***Roundtable on large language models and scientific research*, July 2023**

The Royal Society convened a roundtable on opportunities and risks of using LLMs in scientific research in which Professor Andrew Blake FRS and Gary Marcus provided opening remarks. The roundtable on the use of Large Language Models (LLMs) in scientific research presented both the positive potential and challenges associated with their integration. Participants stressed the importance of developing strategies to mitigate risks and ensuring a balanced approach to the integration of LLMs in research as crucial for the responsible advancement of AI technologies.

Name	Organisation
Alison Noble FRS	University of Oxford
Alistair Nolan	Organisation for Economic Co-operation and Development (OECD)
Andres Guadamuz	University of Sussex
Andrew Blake FRS	Scientific advisor and AI consultant, University of Cambridge
Anthony Cohn	University of Leeds; The Alan Turing Institute
Atoosa Kasirzadeh	University of Edinburgh
Denis Newman-Griffis	University of Sheffield
Edward Tian	GPT Zero
Gabe Gomes	Carnegie Mellon University
Gary Marcus	New York University
Hannah Kirk	University of Oxford; The Alan Turing Institute
Jakob Mökander	Oxford Internet Institute, University of Oxford
James Hetherington	University College London
Jeff Dalton	University of Glasgow
Jessica Montgomery	University of Cambridge
Johan Ordish	Medicines and Healthcare products Regulatory Agency (MHRA)
Matthias Rillig	Freie Universität Berlin
Michael Osborne	University of Oxford; Mind Foundry
Michael Woolridge	University of Oxford
Phil Blunsom	University of Oxford
Samuel Kaski	University of Manchester; Aalto University
Seth Baum	Global Catastrophic Risk Institute
Shreya Rajpal	GuardrailsAI
Yarin Gal	University of Oxford

***Workshop on horizon scanning AI safety risks across scientific disciplines*, October 2023**
Ahead of the Global AI Safety Summit, being organised by the UK Government, the Royal Society will be hosting an official pre-Summit workshop in partnership with the Department for Science, Innovation and Technology. The event brought together senior scientists from academia and industry to horizon-scan the risks associated with AI across scientific disciplines. (See Royal Society website for more information).

Name	Organisation
Alessandro Abate	University of Oxford
Steven Abel	Durham University
Paul Beasley	Siemens
Craig Butts	University of Bristol
Viscount Camrose	House of Lords, DSIT
Sarah Chan	University of Edinburgh
Linjiang Chen	University of Birmingham
Lee Cronin	University of Glasgow
Gwenetta Curry	University of Edinburgh
Christl Donnelly FRS	Imperial College London
Peter Falkingham	Liverpool John Moores University
Tom Fiddian	Innovate UK
Anthony Finkelstein	City, University of London
Michael Fisher	University of Manchester
Jacques Fleuriot	University of Edinburgh
Ben Glocker	Imperial College London
Julia Gog	University of Cambridge
Seraphina Goldfarb-Tarrant	Cohere
Sabine Hauert	University of Bristol
Cathy Holloway	University College London
Caroline Jay	University of Manchester
Alexander Kasprzyk	University of Nottingham
Frank Kelly FRS	Imperial College London
Rohan Kemp	Department for Science, Innovation and Technology
Georgia Keyworth	Department for Science, Innovation and Technology
Ottoline Leyser	UK Research and Innovation
Richard Mallah	Future of Life Institute
Carsten Maple	University of Warwick
Alexandru Marcoci	Centre for the Study of Existential Risk, University of Cambridge

Continued

Name	Organisation
Chris Martin	Department for Science, Innovation and Technology
Emran Mian	Department for Science, Innovation and Technology
Daniel Mortlock	Imperial College London
Gina Neff	University of Oxford
Cassidy Nelson	Centre for Long Term Resilience
Thomas Nowotny	University of Sussex
Yannis Pandis	Pfizer
Nathalie Pettorelli	Zoological Society of London
Reza Razavi	King's College London
Yvonne Rogers FRS	University College London
Sophie Rose	Centre for Long Term Resilience
Stuart Russell	UC Berkeley, Future of Life Institute
Abigail Sellen FRS	Microsoft Research Cambridge
Rossi Setchi	Cardiff University
Nigel Shadbolt FRS	University of Oxford
Shaarad Sharma	Government Office for Science
Karen Tingay	Office for Statistics Regulation
Daniel Tor	Department for Science, Innovation and Technology
Mihaela van der Schaar	University of Cambridge
Mark Wilkinson	University of Sheffield

***Study on red teaming LLM's for resilience to scientific disinformation*, October 2023**

Ahead of the Global AI Safety Summit, being organised by the UK Government, the Royal Society and Humane Intelligence brought together 40 postgraduate students in health and climate sciences to scrutinise how potential vulnerabilities in LLMs (Meta's Llama 2) could enable the generation and spread of scientific misinformation (See Royal Society website for more information)[459].

459 The Royal Society. 2024 Red teaming large language models (LLMs) for resilience to scientific disinformation. See https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/ (accessed 7 May 2024)

APPENDIX 3

Acknowledgements

Working Group members

The members of the Working Group involved in this report are listed below. Members acted in an individual and not a representative capacity and declared any potential conflicts of interest. Members contributed to the project based on their own expertise and good judgement.

Chair
Professor Alison Noble CBE FREng FRS, Foreign Secretary of the Royal Society, and Technikos Professor of Biomedical Engineering, University of Oxford.

Members
Professor Paul Beasley, Head of Research and Development, Siemens.
Professor Peter Dayan FRS, Director, Max Plack 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.
Dr Kelly Vere, Director of Technical Strategy, University of Nottingham

Royal Society staff

Royal Society secretariat
Denisse Albornoz, Senior Policy Adviser and Project Lead
Eva Blum-Dumontet, Senior Policy Adviser (until July 2023)
Areeq Chowdhury, Head of Policy, Data and Digital Technologies
Nicole Mwananshiku, Policy Adviser
Dr Kyle Bennett, Fast Stream placement
Rebecca Conybeare, UKRI placement
Caroline Gehin, UKRI placement

Reviewers

This report has been reviewed by expert readers and by an independent Panel of experts, before being approved by Officers of the Royal Society. The Review Panel members were not asked to endorse the conclusions or recommendations of the report, but to act as independent referees of its technical content and presentation. Panel members acted in a personal and not a representative capacity. The Royal Society gratefully acknowledges the contribution of the reviewers.

Reviewers
Dr Yoshua Bengio FRS, Professor at University of Montreal and Scientific Director of MILA
Ruhi Chitre, Ezra Clark, Tiffany Straza and Ana Persic (Natural Sciences Sector); and Irakli Khodeli (Social and Human Sciences Sector), UNESCO
Dr Rumman Chowdhury, CEO and Founder of Humane Intelligence. 2024 US Science Envoy.
Professor Tony Hey FREng, Honorary Senior Data Scientist at Rutherford Appleton Laboratory. Co-author of *Artificial Intelligence For Science: A Deep Learning Revolution*

The Royal Society is a self-governing Fellowship of many of the world's most distinguished scientists drawn from all areas of science, engineering, and medicine. The Society's fundamental purpose, as it has been since its foundation in 1660, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity.

The Society's strategic priorities emphasise its commitment to the highest quality science, to curiosity-driven research, and to the development and use of science for the benefit of society. These priorities are:

- The Fellowship, Foreign Membership and beyond
- Influencing
- Research system and culture
- Science and society
- Corporate and governance

For further information
The Royal Society
6 – 9 Carlton House Terrace
London SW1Y 5AG

T +44 20 7451 2500
W royalsociety.org

Registered Charity No 207043

9 781782 527121

ISBN: 978-1-78252-712-1
Issued: October 2024 DES8836_5