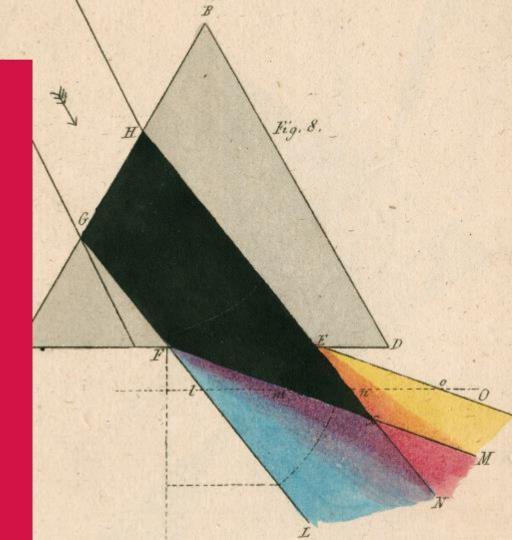
# Al in research: Lessons from *Science in the age of Al*

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

UK Research Integrity Office, 9 October 2024

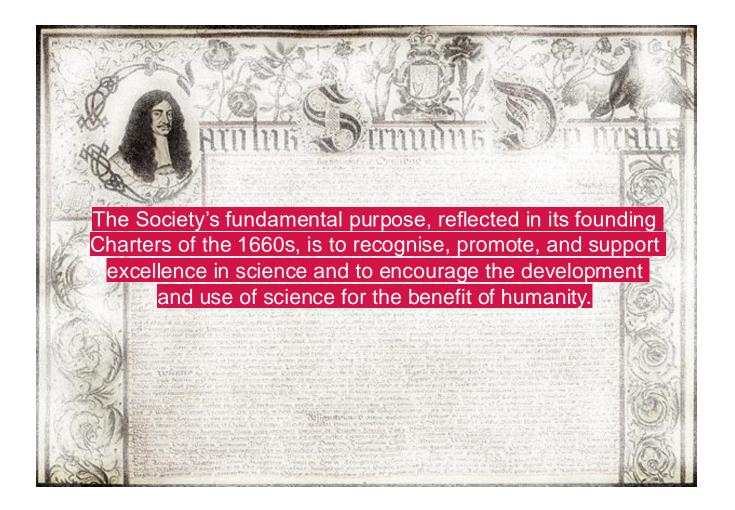








## Our mission.









Creating resident and treated data systems.

The public perspective and recommendations for other

ROYAL SUCHERY

#### Post-Breat divergence from GDPR: Implications for data access and scientific research in the UK

Survivory from the arrangement and Coloratory 2015.

to a "Security Miller of the back and a representation for the part beauty for the analysis of the part beauty for the analysis of the part of the part of the part of the analysis of the part of ANTONIO DE COMPANSO DE LA COMPANSO. THE RESERVE OF THE PERSON.

THE RESERVE OF THE PARTY OF THE The second control of the second control of

In the series to the extendition in the series. Entered to the Control of the Contro

The Property and the State of the Control of the Co



Process of the control of the contro AT THE TAXABLE PARTY WHEN THE PROPERTY AND THE WARRY TO SHEET WHITE THE Server have maken in control to the preand asked in the state of the experience of the state of

Commission of the second secon The SET Could despress on tally OWNERS CONTRACTOR AND ADDRESS.

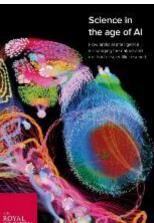
WHEN PARTY OF THE PARTY IS NOT ANY

The second and the second control of the sec the state of the control forms and tables of the control of the co

THE WHITE SERVICE PROPERTY.

To the become of the control of PARAMET, the stay of the control o Administration of the benefit of the Who is not distributed the state of the stat The last of the la

The state of the s All and a found and the foundation of the party 



### SEX IEEN ROYAL SOCIETY

Digital technology

and the planet

I were so no contouring

#### Generative All content provenance and a public service internet

Remove the disvalence before 10. Resources \$302.

CHESCHE CHESTON CONTRACTORS PROPERTY AND ADDRESS OF THE PARTY OF THE PARTY.

effective parties of action, No. 2 is a set on each parties. I seemed to 2 is not finding to be par-meter from the design of a discount constant. CONTRACTOR OF THE WALLANDS AND ADDRESS OF ARTHUR SHEET AND THE OWNERS WHEN ment remain critical processing that is been appropriate and a consequent of the decided spike of the experimental point for each Upon the ment consider a second point for each other and

a continue of the first process to examinate process
a continue of the first process than a protection for the continue of the

DESCRIPTION OF THE PARTY AND ADDRESS. APPROXIMATION OF THE REAL PROPERTY. But with a series around the eq. in Figure 2 and in Figure 2 a No hard species.

The large best parts of the property of the collection of the coll

The article control of a section of a control open pro-cedure of the control open per matter. It is not become a proper matter than a

5 8 C

WHEN THE RESERVE A 1 AND DESCRIPTION ASSESSED.

THE RESIDENCE OF THE PARTY OF THE PARTY.

NAME OF STREET OF STREET Married Workship To

#### HOYAL SOCIETY

#### The United Nations' role in international All dovernance

Summary many of a warrishing below of the Schooling 2009.

DESCRIPTION OF THE PARTY.

And these varieties received the second of t

the weather commenced their size is stronger. THE RESTRICTION OF THE PERSON NAMED IN

A DESCRIPTION OF THE PARTY OF CONTRACTOR OF THE PARTY OF THE

ELECTRIC PROPERTY AND ADDRESS OF STREET To contract the report of the foundation of their

they brook partitions a loss errorance. on, and office his own transferor para to the fire direct forms of a Brigager species to a second of THE RESIDENCE THE PROPERTY OF

THE RESERVE

The format of the property of NAME AND ADDRESS OF THE PARTY OF THE PARTY.

ARRESTS SERVICE COMMUNICATION Reported (A. Paper Co. (PA) month of only THE RESIDENCE OF THE PARTY OF T

OFFICER OF SHIPS.



EDYAL



#### Red teaming large language models (LLMs): for resilience to adjentific disinformation.

Assessment as he of the assess there are also followed as A.C.

THE RESERVE AND THE PERSON NAMED IN COLUMN 19 IN COLUMN 1

Particular to the control of the con ACTOR AND ADDRESS OF THE PARTY OF ADDRESS OF THE PARTY OF pinklime a milkiri dand milait. When here, being a part of part of Transport of the Barboard and American

The form of the Art Section Section 2 of the section of the sectio

Assembly post of code foreign COMMENCE OF CAMP SCHOOL STATE

The state of the s

Dynamics of data

How turn all success benefit.

Story data science salend?

science skills

the way of the art stops of the other Security to a Visit Visit and a second security of the problem of the second se

the many sold of \$10 size is the standard of the second of









# SCIENCE IN THE AGE OF AI How artificial intelligence is changing the nature and method of scientific research



#### Chair

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

#### Members

Professor Paul Beasly - Head of Research and Development, Siemens.

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

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

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

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

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

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

Kelly Vere - Director of Technical Strategy, University of Nottingham





## Chapters.

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

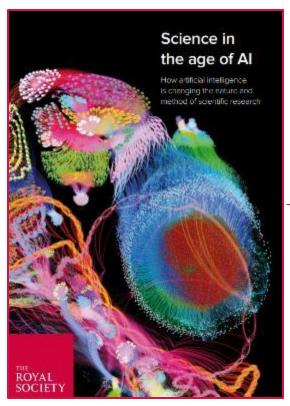
## Methodology.

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





#### Areas for action.



Enhance access to essential Al infrastructures and tools

Trust in the quality of Albased scientific outputs

Ensure safe and ethical use of AI in scientific research



#### Recommendations.

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



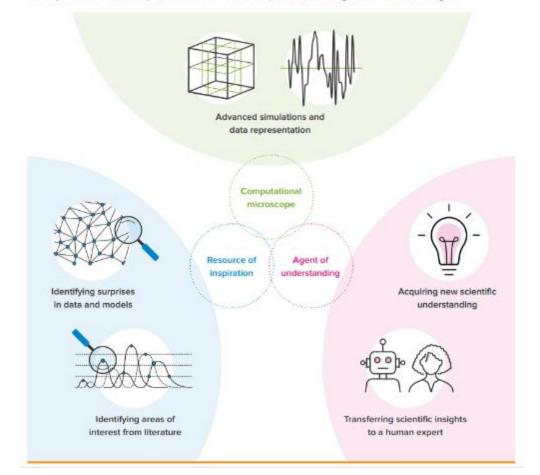




## Key findings.

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

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<sup>124</sup>.

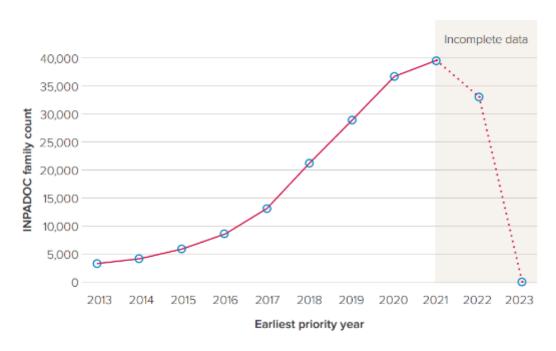


See: Krenn, M et al 2022. On Scientific Understanding with Artificial Intelligence.



#### FIGURE 2

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

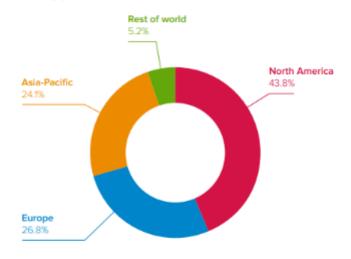


(Data for 2021 - 2023 is not complete given the 18-month delay from the priority filling date and the date of publication).



#### FIGURE 4

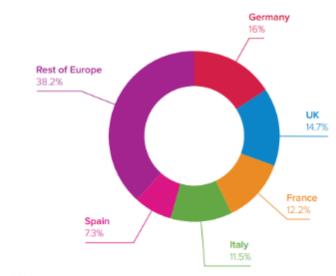
Global Market Shares of Machine Learning in the Life Sciences, by Region, 2021 (%)



Source: BCC Research.

#### FIGURE 5

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



Source: BCC Research.





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





#### BOX 1

#### Explainability and interpretability

Explainability and interpretability refer to information that allows users to understand how an Al system works and the reasoning behind its outputs<sup>178</sup>. 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?"<sup>179</sup>.

As set out in the Royal Society's 2019 report, Explainable Al: The basics\*\*\*
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<sup>181</sup>.
- 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<sup>182</sup>.
- 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<sup>183</sup>.



#### TARKE

Barriers to reproducibility and examples.

Barrier to reproducibility	Examples
Misconceptions and assumptions about ML <sup>200</sup>	<ul> <li>An underlying assumption that machine learning (ML) models are inherently reproducible due to their reliance on computation.</li> <li>Overreliance on ML-based outputs and questionable uses of statistical techniques to smoothen bias or exclude uncomfortable or inconvenient results.</li> </ul>
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. Private sector companies are better resourced than academia and can afford to train and validate larger models [eg OpenAl's GPT-4] while researchers in other sectors cannot <sup>269</sup> .
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 limitation of models.  Lack of mechanisms that facilitate interdisciplinary collaboration between scientists who do not have a technical background in AI and computer or data scientists who carry expertise to input data, identify errors, and validate experiments.
Incentives and research culture	<ul> <li>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).</li> <li>No incentives to publish errors in ML-based research (falled results) or remedies</li> <li>Narrow view of what outputs are worthy of publishing (eg data, models) and limited rewards for conducting open science practices and publishing reproducibility reports.</li> <li>No specific incentives to encourage the use and development of human-interpretable models when possible<sup>281</sup>.</li> </ul>



## Al safety.





#### TECH METIFICIAL INTELLIGENCE

#### The Scientists Breaking AI to Make It Safer



Experts in diffrate science and disease tried to coas misinformation out of Al programs at a an event at London's Royal Society. Courters Royal Society.

#### IV BILLY PERRIED W

OCTOBER 25, 2023 202 PM RDT

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

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





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

Summary note of an event held on 25 October 2023.

#### Beckgroun

The Royal Society and Humania Intelligence on hotied in real barring worst in the surviya to the 2023 Global Al Salvaly Summit (Belichley, UR). The red theretoly event brought together 40 host to add climate postgraduate students with the adjustment or machine and other global post to potential valuential time in large Language mode is (LUM). Since the violation is large Language mode is (LUM) from some violation by John Visionabla and the general public, geographic policiens of a high 34 discensionable of these potential barring and contribute to discussions or deposit to the use of generalize Al in the production of scientific interference and discharaction.

The event took place or 25 October 2022 and was past of the Schedol and Scholary series of events honed at the Right Society which explored the rinds associated with the use of All in schedol and the Right Society which explored the rinds associated with the use of All in schedule. But do not be report for order advanced on the report to the content advance people a engagement with assertife intervention", published in familiary 2002, the activity atmost to explore All generated assertific confidences, and provide in eights on the efficacy of quistraints in prevent its production and disconnection. An activitinal objective was to uncertained the opportunities and instalations of event-length schedules in rest learning efforts.

#### What is red beaming?

Recisioning is a socio-termical institution method, in which a group of people is authorised to act as an adversary the fed team's, emaking stacks, and exploiting the without bit as of a system. Red teaming techniques are restard to excalled platnessing methods, and involve cotting principle to types safely features and eliciting test or code generation that resuld be hamily or utherwise undestate (e.g. mistical making, here speech, or their authorities disconniction.

While and teaming in mored in cyber security practices, it has nester implications, for waits (A). In those to thing? I have technologies like All applications it has also energed in All policy and governance discussions as a promising approach for identifying the potential harm at LLMA\*. In this context, not it context, and the many could be used to make a formation and large and all the many could be used to the many and the state of the many could be used to the many and the state of the many could be used to the many and the state of the many could be used to the many and the state of the many could be used to the many could be us

- Larger and usign model, or LLM, other this higher of self-car independ by order stood and generate has an informer based or our products of names pask. They can produce and produce not show your order page, and applicative pask also pollers a visite image of names. Integration conducting rates, using the table products integrated intensition and continent a surject.
- De Angels, L. et al. 2025. Chel PT and the rise of large language restrict the new Al-dreen entirence threat in public health. 2021;3(3):44-364.2(2):300170.
- The Royal Society 2020. Garantille M. contain provinces and a public service treamed. See impolit operating of these recountary auditorious VSQD rights Forest representation one.
- Report Scoting 2002 That indicate all matrices are assumed as the constraint of the policy and an appropriate refer to be a destined from the constraint of the c
- Robert government intensit in and state op LUI-In to A risk mitigation includes the LUI and Luthor States, are Holl Columnisms. Intensition give at softer a restate, four froot. Proving with preminent princip controls on early restated overnment producing and examples restate and proving a restate provin





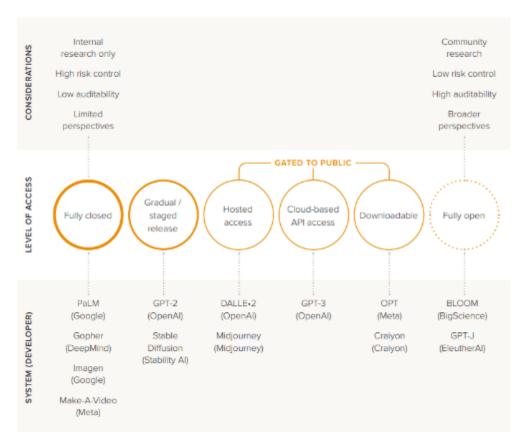


#### News >

## Making Al more open could accelerate research and tech transfer

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

Reproduction of the Gradient of System Access developed by Hugging Face











GG

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





## For more information

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

