https://doi.org/10.1038/s42254-023-00581-4

Science in the age of large language models

Abeba Birhane, Atoosa Kasirzadeh, David Leslie & Sandra Wachter

Rapid advances in the capabilities of large language models and the broad accessibility of tools powered by this technology have led to both excitement and concern regarding their use in science. Four experts in artificial intelligence ethics and policy discuss potential risks and call for careful consideration and responsible usage to ensure that good scientific practices and trust in science are not compromised.

Large language models (LLMs) are deep learning models with a huge number of parameters trained in an unsupervised way on large volumes of text. LLMs started to emerge around 2018 and since then there has been a sharp increase in the number of parameters and capabilities (for example, GPT-4 has over 100 trillion parameters and can process both text and images). Discussions about the use and misuse of this technology in science erupted in late 2022, prompted by the sudden widespread access to LLM tools that can generate and edit scientific text or can answer scientific questions. Some of the open questions fuelling these conversations are summarized in Box 1.

What are the wider concerns?

Abeba Birhane: In a matter of months, LLMs have come to captivate the scientific community, general public, journalists and legislators. These systems are often presented as game-changers that will radically affect our lives from the way we search for information to the way we create art and do science. As hype around the capabilities of these systems continues to grow, many claims are made without evidence; the burden of disproving these claims is put on critiques. Despite the concrete negative consequences of these systems on actual people[1] — often on those at the margins of society — questions of responsibility, accountability, exploited labour and otherwise critical inquiries rarely accompany discussion of LLMs. Instead, discussions are dominated by abstract and hypothetical speculations around their intelligence, consciousness, moral status and capability for understanding, all at the cost of questions of responsibility, underlying exploited labour and uneven distribution of harm and benefit from these systems.

Sandra Wachter: Generative AI (GenAI, deep learning models that can output data beyond text, such as images or audio), more broadly, is a potentially very disruptive technology that could impact many areas such as education, media, art and scientific research. The disruption of both the production and consumption of science and research is particularly concerning because domain expertise is necessary to detect when GenAI has 'hallucinated' or invented falsehoods and confidently passed them off as the truth.

Disruptive technologies have always inspired great hopes and fears. The printing press was feared to lead to the moral erosion of society, fast moving automobiles were assumed to harm internal organs of people and the telephone was said to destroy family values. Many of these fears were ultimately unfounded. But other dangers did materialize that were not even on the radar of developers, scholars and policymakers at the time, such as the significant impact of personal automobiles on the environment. Reliably predicting the social and economic impacts, risks and development pathway of disruptive technologies is difficult. This is not to say that we should stop horizon scanning, but rather that we need to periodically re-evaluate the risks and benefits of technologies.

Among these risks, the environmental impact of these technologies urgently needs to be addressed. Regardless of their utility, we need to keep in mind that they have a significant carbon footprint[2]. As opposed to when the automobile first appeared, we now know the environmental costs society is forced to bear. As scientists, and as a society, we must not look away from how the use of artificial intelligence (AI) technologies can exacerbate the climate crisis.

What are the specific concerns for science?

David Leslie: LLMs, and more broadly foundation models and GenAI, will undoubtedly play a significant role in the future of scientific discovery. Researchers, however, must proceed with caution, engaging the affordances provided by these technologies with the same kinds of epistemic humility, deflationary scepticism and disciplined adherence to the scientific method that have functioned as preconditions of modern scientific advancement since the dawn of the seventeenth-century Baconian and Newtonian revolutions. Amidst the hype surrounding LLMs, scientists must acknowledge the social and interpretative character of scientific discovery and manage expectations regarding the contributions of LLMs to the advancement of scientific understanding.

LLMs generate predictions of the 'statistically likely continuations of word sequences'[3] based on brute-force iterative training on massive corpuses of digital text data. As sequence predictors, these models draw on the underlying statistical distribution of previously generated text to stitch together vectorized symbol strings based on the probabilities of their co-occurrence[4]. They therefore lack the communicatively embodied and relational functionings that are a prerequisite of scientific meaning-making, in the barest sense. These systems do not 'inhabit' the lived reality in which speaking and interacting members of the human community together build and reproduce a common world of shared experience, using the agency of language to convey intention, to assess and establish truth through the exchange of reasons and to cope with the myriad problems of existence. In this way, LLMs, foundation models and GenAI technologies lack the basic capacities for intersubjectivity, semantics and ontology that are preconditions for the kind of collaborative world-making that allows scientists to theorize, understand, innovate and discover. Despite their impressive feats of rhetorical prowess, systems such as ChatGPT can neither navigate an evolving space of scientific reasons nor partake in the trials and triumphs of scientific meaning-making. Their subsidiary role in scientific discovery should hence be understood taking this limitation into account.

Atoosa Kasirzadeh: I point to three significant concerns regarding the use of LLMs in scientific contexts. First, LLMs may not capture nuanced

value judgements implicit in scientific writings[5]. Although LLMs seem to provide useful general summaries of some scientific texts, for example, it is less clear whether they can capture the uncertainties, limitations and nuances of research that are obvious to the human scientist. Relying solely on LLMs for writing scientific summaries can result in oversimplified texts that overlook crucial value judgements and lead to misinterpretations of study results. We should, therefore, proceed with caution when using LLMs for scientific summarization. Additional work is needed to ensure that LLMs accurately communicate the value judgements underlying scientific practice. This work should include designing appropriate evaluation benchmarks to assess the accuracy of LLMs in communicating these value judgements.

Second, LLMs have been known to generate non-existent and false content — a phenomenon that has been dubbed 'hallucination'. For instance, Meta's Galactica, an LLM that was initially designed to reason about scientific knowledge, was reported to exhibit significant flaws such as reproducing biases and presenting falsehoods as facts[6] and was shut down after only 3 days of public API access. Therefore, overreliance on LLMs for tasks such as writing literature reviews should be avoided. Or at least the output should be very carefully fact-checked.

Third, the use of LLMs in the peer-review process can endanger trust in it. LLMs used for writing peer-review reports run the risk of misinterpreting the submitted scientific article, be it by a loss of crucial information or by a hallucination in the aforementioned sense. And whereas one can hold human reviewers responsible, it is a nontrivial question how to hold LLMs responsible — in part owing to their opaque nature. It seems like a responsibility gap is lurking here.

Who bears the responsibility?

AB: As we rush to deploy LLMs into scientific practices, it is important to remember that science is a human enterprise and LLMs are tools — albeit impressive at predicting the next word in a sequence based on previously 'seen' words — with limitations such as brittleness (susceptibility to catastrophic failure), unreliability and the fabrication of seemingly 'scientific' nonsense. Even if these limitations can, by some miracle, be solved, it would be a grave error to treat LLMs as scientists that can produce science. Knowledge implies responsibility and is never detached from the scientist that produces it. Science never emerges in a historical, social or cultural vacuum and builds on a vast edifice of well-established knowledge. We embark on a scientific journey to build on this edifice, to react and to debunk it, in anticipation of responses and reactions. We take responsibility for our work and defend it when criticized or retract it when proven wrong. What is conceived as science can be dependent on ideologies of the time. For example, at its peak during the early nineteenth century, eugenics was mainstream science. Most importantly, as science is never done from a 'view from nowhere', our questions, methodologies, analysis and interpretations of our findings are influenced by our interests, motivations, objectives and perspectives. LLMs, as tools, have none of these. As tools, LLMs, with close and constant vetting by the scientist, can aid scientific creativity and writing[7]. However, to conceive of LLMs as scientists or authors themselves is to misunderstand both science and LLMs and to evade responsibility and accountability.

What should scientists do?

SW: We are currently at a crucial point with GenAI. Its possibilities seem limitless, and yet we are still early enough in its lifecycle to transform its future pathway. Science is fast paced and highly competitive. The pressure to publish can be overwhelming. A technology that can save time in conducting research and increasing output can be very tempting. But if GenAI is used automatically and without critical oversight, it may fundamentally undermine the foundations of 'good' science.

At this stage, we need to think about how to responsibly integrate GenAI into science. Scientists have an ethical responsibility to society to produce knowledge that follows the highest possible standards. Climate change and COVID-19 are just two examples of the overwhelming importance of reliable science for driving policy and societal action. Researchers need to collaborate with journals, publishers, conference organizers, the press and the wider scientific community to develop best practices, standards and detection methods to ensure that the benefits of GenAI can be realized without fundamentally undermining science and its role in society.

DL: Scientists must view LLMs and GenAI technologies as exploratory tools that bolster responsible, mission-driven and society-led research practices and that support the advancement of scientific discovery and understanding. To paraphrase the words of economist Zvi Griliches[8], the expanding use of these AI technologies in scientific research is the 'discovery of a method of discovery' — the invention of a new set of research tools that support and enable new pathways of insight, innovation and ingenuity in the physical and life sciences.

Starting from such a tool-based understanding, researchers must view the role of these technologies in scientific discovery through a chastening, but non-reductive lens, deploying them as computational vehicles of observation and analysis to probe properties of complex physical and biological systems and patterns in high-dimensional biophysical data that would otherwise be inaccessible to human-scale examination, experiment and inference. But the path to discovery should not be treated in a strictly instrumentalist way; scientists should not see these complex models as mere oracles. Rather, their results and innerworkings should be seen as springboards for scientific reflection and

Author contributions

A.B. is cognitive scientist researching human behaviour, social systems and responsible and ethical AI. She is a Senior Fellow in Trustworthy AI at Mozilla Foundation and an Adjunct Assistant Professor at Trinity College Dublin, Ireland.

A.K. is a philosopher and ethicist of science and emerging technologies, an applied mathematician and an engineer. Currently, she is a tenure-track assistant professor and a Chancellor's Fellow in the Philosophy department and the Director of Research at the Centre for Technomoral Futures in the Futures Institute at the University of Edinburgh. Her recent work is focused on the implications of machine learning, in particular large language models and other models for science, society and humanity.

S.W. is Professor of Technology and Regulation at the Oxford Internet Institute at the University of Oxford where she researches the legal and ethical implications of AI, Big Data and robotics as well as Internet and platform regulation. At the OII, she leads and coordinates the Governance of Emerging Technologies (GET) Research Programme that investigates legal, ethical and technical aspects of AI, machine learning and other emerging technologies.

D.L. is Professor of Ethics, Technology and Society at Queen Mary University of London and the Director of Ethics and Responsible Innovation Research at The Alan Turing Institute. He is a philosopher and social theorist, whose research focuses on the ethics of emerging technologies, AI governance, data justice and the social and ethical impacts of AI, machine learning and data-driven innovations.

creativity that can play a constituent role in guiding the broader socially embodied pursuit of the expansion and refinement of scientific understanding[9].

In addition, the AI-generated outputs and the insights of these models must be regarded as both interpreter-dependent and theory-laden. The construction and deployment of LLMs and GenAI tools and their application in scientific exploration must be seen as interpretive accomplishments that are embedded in what philosophers of science from have called 'contexts of discovery'[10,11]. These are contexts of scientific sense-making that involve real-life processes of communication carried out cooperatively by members of an unbounded human community of inquiry, interpretation and reason-giving.

AK: Until more robust and reliable safeguards are in place, the scientific community should take a timely and firm stance to avoid any overreliance on LLMs and to foster practices of responsible science in the age of LLMs. Otherwise, the risk is to jeopardize the credibility of scientific knowledge. An initial step towards this is to try to design LLM policies in a realistic way; for example, to identify and ban papers that primarily rely on LLMs, a policy already adopted at the International Conference on Machine Learning (ICML) 2023 and likely to be enforced widely. However, identifying LLM-generated text is challenging, and the development of accurate detection tools is an ongoing area of research. Recent studies have raised concerns about the reliability of these methods in accurately distinguishing between LLM-generated and non-LLM-generated text[12].

In addition, scientists must also be more vocal about the potential negative impacts of this technology on the scientific community. By raising awareness and demanding further research and development of safeguards, the scientific community can actively contribute to the responsible and ethical use of LLMs.

Box 1

Open questions

Accuracy, reliability and accountability

- Hallucination: How can scientists methodically determine when large language models (LLMs) are 'hallucinating' or generating inaccurate and fantastical content? How can scientists best assess and work around these tendencies to generate unreliable or non-factual outputs?
- Responsiveness to change: If LLMs fail to extrapolate effectively when world knowledge changes or data distributions drift over time, how can scientists safeguard their accuracy, reliability and responsiveness to change?
- Sparse phenomena: If LLMs struggle to reliably generate accurate content for infrequent or sparsely studied phenomena, how do scientists draw on LLMs to inform insights about anomalies, new discoveries or unprecedented observations?
- Research integrity: What is plagiarism and authorial misrepresentation in the age of LLMs? How should scientists be held accountable for plagiarism and authorial misrepresentation? What checks should be put in place to establish the authenticity of scientific publications?
- Quantifying the degree of LLMs assistance in writing: What is acceptable and what is not?
- Accountability: Who is responsible for the integrity of scientific research and the content of scientific papers aided by LLMs? Who is accountable?

Explainability, missingness and bias

- Opacity: How can opaque LLMs justifiably be integrated into the scientific method?
- Explainability: How can the original sources be traced back? How can scientists, who draw on opaque LLMs, clarify the intended meaning or nuances of the texts based on which such models render their outputs? Does a lack of interpretability undermine the justifiability of relying on inferences drawn from LLMs?
- Missingness: If scientific papers represent the final product of a research process rather than a full picture of the complex choices, practices and contexts that underlie the research (that is not all research is documented, in particular failures and negative results), how can the inferences generated by LLMs (which only process the information scientific articles, textbooks, websites and so on) account for the missingness that derives from the limitations of such a 'tip-of-the-iceberg' view?
- Selection: How can LLMs account for outdated or incorrect knowledge in the published literature?
- Bias: How can potential biases in the training data sets of LLMs — and other social, statistical and cognitive biases that may arise in their design, development and deployment — be most effectively assessed? How will LLMs enhance existing and introduce new biases or help remove existing ones?

Scientific ingenuity and discovery

- Paradigm shifts: How can LLMs accommodate future 'paradigm shifts' in scientific understanding? Could LLMs (which generate insights by identifying patterns emergent from past research — potentially engendering paradigm lock-in and stifling novelty) function to tamp down possibilities for new scientific directions?
- Outliers: Will outliers (radical new ideas, unconventional views and unusual writing styles) be lost, overlooked or averaged out?
- Scientific creativity: What is the role of the scientist in the age of LLMs? What is the role of scientific creativity?
- Deskilling: Will overreliance on LLMs to produce arguments and text risk diminishing or weakening the writing and critical thinking skills and insight of researchers?

Science assessment and peer review

- Assessing quality: How do we assess high-quality science in the age of LLMs? What role should the values of reproducibility/replicability and transparency play?
- Ethos of science: How do we trust science in the age of LLMs? How, if at all, do the values of objectivity, rigour and accountability change with the scaled integration of LLMs into scientific practices?

This includes promoting interdisciplinary collaboration and sharing knowledge about the potential risks and benefits of LLMs in various fields.

It is important for the scientific community to closely monitor these developments and to urge AI research laboratories, such as OpenAI, to prioritize research on more reliable detectors. Furthermore, it is crucial that the scientific community continues to closely follow the development and use of LLMs and adapts their policies and practices in consultation with AI ethics and safety experts, to ensure that the use of LLMs enhances, rather than undermines, the rigor and reproducibility of scientific research. Finally, the scientific community must encourage more interdisciplinary discussions with experts from academia and industry to navigate the implications of LLMs for scientific knowledge.

Abeba Birhane[1] ✉, **Atoosa Kasirzadeh**[2,3] ✉, **David Leslie**[3,4] ✉ **& Sandra Wachter**[5] ✉
[1]Mozilla Foundation and Trinity College Dublin, Dublin, Ireland. [2]The University of Edinburgh, Edinburgh, UK. [3]The Alan Turing Institute, London, UK. [4]Queen Mary University of London, London, UK. [5]University of Oxford, Oxford, UK.
✉e-mail: adbirhane@gmail.com; atoosa.kasirzadeh@ed.ac.uk; dleslie@turing.ac.uk; sandra.wachter@oii.ox.ac.uk

Published online: 26 April 2023

References

1. Weidinger, L. et al. Taxonomy of risks posed by language models. in *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency* 214–229 (ACM, 2022).
2. Bender, E. et al. On the dangers of stochastic parrots: can language models be too big? in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (ACM, 2021).
3. Shanahan, M. Talking about large language models. Preprint at https://doi.org/10.48550/arXiv.2212.03551 (2022).
4. Bender, E. & Koller, A. 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 (ACL, 2020).
5. Kasirzadeh, A. & Gabriel, I. In conversation with artificial intelligence: aligning language models with human values. *Philos. Technol.* **36**, 27 (2023).
6. Heaven, W. D. Why Meta's latest large language model survived only three days online, MIT Technology Review. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ (2023).
7. Owens, B. How Nature readers are using ChatGPT. *Nature* https://www.nature.com/articles/d41586-023-00500-8 (20 February 2023).
8. Griliches, Z. Hybrid corn: an exploration in the economics of technological change. *Econometrica* **25**, 501–522 (1957).
9. Krenn, M. et al. On scientific understanding with artificial intelligence. *Nat. Rev. Phys.* **4**, 761–769 (2022).
10. Reichenbach, H. Experience and prediction. An analysis of the foundations and the structure of knowledge. *J. Philos.* **35**, 270 (1938).
11. Kuhn, T. *The Structure of Scientific Revolutions* (University of Chicago Press, 2012).
12. Sadasivan, V. S. et al. Can AI-generated text be reliably detected? Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.11156 (2023).

Acknowledgements

The work of S.W. is supported through research funding provided by the Wellcome Trust (grant nr 223765/Z/21/Z), Sloan Foundation (grant no. G-2021-16779), the Department of Health and Social Care (via the AI Lab at NHSx) and Luminate Group to support the Trustworthiness Auditing for AI project and Governance of Emerging Technologies research programme at the Oxford Internet Institute, University of Oxford.

Competing interests

The authors declare no competing interests.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.