PNAS EDITORIAL

Protecting scientific integrity in an age of generative AI

Wolfgang Blau[a], Vinton G. Cerf[b], Juan Enriquez[c], Joseph S. Francisco[d], Urs Gasser[e], Mary L. Gray[f,g], Mark Greaves[h], Barbara J. Grosz[i], Kathleen Hall Jamieson[j], Gerald H. Haug[k], John L. Hennessy[l], Eric Horvitz[m], David I. Kaiser[n], Alex John London[o], Robin Lovell-Badge[p], Marcia K. McNutt[q,1], Martha Minow[r], Tom M. Mitchell[s], Susan Ness[j], Shobita Parthasarathy[t], Saul Perlmutter[u,v], William H. Press[w], Jeannette M. Wing[x], and Michael Witherell[y]

Revolutionary advances in AI have brought us to a transformative moment for science. AI is accelerating scientific discoveries and analyses. At the same time, its tools and processes challenge core norms and values in the conduct of science, including accountability, transparency, replicability, and human responsibility (1–3). These difficulties are particularly apparent in recent advances with *generative AI*. Future innovations with AI may mitigate some of these or raise new concerns and challenges.

With scientific integrity and responsibility in mind, the National Academy of Sciences, the Annenberg Public Policy Center of the University of Pennsylvania, and the Annenberg Foundation Trust at Sunnylands recently convened an interdisciplinary panel of experts with experience in academia, industry, and government to explore rising challenges posed by the use of AI in research and to chart a path forward for the scientific community. The panel included experts in behavioral and social sciences, ethics, biology, physics, chemistry, mathematics, and computer science, as well as leaders in higher education, law, governance, and science publishing and communication. Discussions were informed by commissioned papers detailing the development and current state of AI technologies; the potential effects of AI advances on equality, justice, and research ethics; emerging governance issues; and lessons that can be learned from past instances where the scientific community addressed new technologies with significant societal implications (4–9).

Generative AI systems are constructed with computational procedures that learn from large bodies of human-authored and curated text, imagery, and analyses, including expansive collections of scientific literature. The systems are used to perform multiple operations, such as problem-solving, data analysis, interpretation of textual and visual content, and the generation of text, images, and other forms of data. In response to prompts and other directives, the systems can provide users with coherent text, compelling imagery, and analyses, while also possessing the capability to generate novel syntheses and ideas that push the expected boundaries of automated content creation.

Generative AI's power to interact with scientists in a natural manner, to perform unprecedented types of problem-solving, and to generate novel ideas and content poses challenges to the long-held values and integrity of scientific endeavors. These challenges make it more difficult for scientists, the larger research community, and the public to 1) understand and confirm the veracity of generated content, reviews, and analyses; 2) maintain accurate attribution of machine- versus human-authored analyses and information; 3) ensure transparency and disclosure of uses of AI in producing research results or textual analyses; 4) enable the replication of studies and analyses; and 5) identify and mitigate biases and inequities introduced by AI algorithms and training data.

Five Principles of Human Accountability and Responsibility

To protect the integrity of science in the age of generative AI, we call upon the scientific community to remain steadfast in honoring the guiding norms and values of science. We endorse recommendations from a recent National Academies report that explores ethical issues in computing research and promoting responsible practices through education and training (3). We also reaffirm the findings of earlier work performed by the National Academies on responsible automated research workflows, which called for human review of algorithms, the need for transparency and reproducibility, and efforts to uncover and address bias (10).

Building upon the prior studies, we urge the scientific community to focus sustained attention on five principles of human accountability and responsibility for scientific efforts that employ AI:

1. **Transparent disclosure and attribution**
   Scientists should clearly disclose the use of generative AI in research, including the specific tools, algorithms, and

Author affiliations: [a]Brunswick Group, London WC2A 3ED, United Kingdom; [b]Google Global Networking, Google LLC, Reston, VA 20190; [c]Excel Venture Management, Boston, MA 02116; [d]Department of Earth and Environmental Science and Department of Chemistry, University of Pennsylvania, Philadelphia, PA 19104; [e]School of Social Sciences and Technology, Technical University of Munich, 80333 München, Germany; [f]Microsoft Research New England Lab, Cambridge, MA 02142; [g]Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN 47408; [h]Schmidt Sciences, Washington, DC 20008; [i]Harvard Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; [j]Annenberg Public Policy Center, University of Pennsylvania, Philadelphia, PA 19104; [k]German National Academy of Science Leopoldina, 06108 Halle, Germany; [l]Knight-Hennessy Scholars, Stanford University, Stanford, CA 94305; [m]Office of the Chief Scientific Officer Microsoft, Redmond, WA 98052; [n]Program in Science, Technology, and Society and Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139; [o]Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213; [p]Laboratory of Stem Cell Biology and Developmental Genetics, The Francis Crick Institute, London NW1 1AT, United Kingdom; [q]National Academy of Sciences, Washington, DC 20001; [r]Harvard University, Cambridge, MA 02138; [s]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213; [t]Science, Technology, and Public Policy Program, University of Michigan, Ann Arbor, MI 48109; [u]Department of Physics, University of California, Berkeley, CA 94720; [v]Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [w]Department of Computer Science, The University of Texas at Austin, Austin, TX 78712; [x]Office of the Executive Vice President for Research and Computer Science Department, Columbia University, New York, NY 10027; and [y]Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Competing interest statement: V.G.C., M.L.G., and E.H. are based at organizations that are engaged with developing and fielding AI technologies. J.E. is a Managing Director of a venture capital firm. Many of his businesses are beginning to use and deploy AI. K.H.J. is Program Director of the Annenberg Foundation Trust at Sunnylands, a co-convener of the retreats whose participants authored the statement. J.L.H. is a Director of Alphabet, which invests in and uses AI in its business. M.K.M. is the President of the National Academy of Sciences. T.M.M. received a grant from Microsoft for computing on their Azure cloud computing facility and owns stock in a variety of AI-related companies, Alphabet, Amazon, Meta, Microsoft, and Nvidia. M.W. works for a research institution that is engaged in developing and applying AI in scientific research. All other authors declare no competing interests.

[1]To whom correspondence may be addressed. Email: naspresident@nas.edu.

Published May 21, 2024.

settings employed; accurately attribute the human and AI sources of information or ideas, distinguishing between the two and acknowledging their respective contributions; and ensure that human expertise and prior literature are appropriately cited, even when machines do not provide such citations in their output.

Model creators and refiners should provide publicly accessible details about models, including the data used to train or refine them; carefully manage and publish information about models and their variants so as to provide scientists with a means of citing the use of particular models with specificity; provide long-term archives of models to enable replication studies; disclose when proper attribution of generated content cannot be provided; and pursue innovations in learning, reasoning, and information retrieval machinery aimed at providing users of those models with the ability to attribute sources and authorship of the data employed in AI-generated content.

2. **Verification of AI-generated content and analyses**
Scientists are accountable for the accuracy of the data, imagery, and inferences that they draw from their uses of generative models. Accountability requires the use of appropriate methods to validate the accuracy and reliability of inferences made by or with the assistance of AI, along with a thorough disclosure of evidence relevant to such inferences. It includes monitoring and testing for biases in AI algorithms and output, with the goal of identifying and correcting biases that could skew research outcomes or interpretations.

Model creators should disclose limitations in the ability of systems to confirm the veracity of any data, text, or images generated by AI. When verification of the truthfulness of generated content is not possible, model output should provide clear, well-calibrated assessments of confidence. Model creators should proactively identify, report, and correct biases in AI algorithms that could skew research outcomes or interpretations.

3. **Documentation of AI-generated data**
Scientists should mark AI-generated or synthetic data, inferences, and imagery with provenance information about the role of AI in their generation, so that it is not mistaken for observations collected in the real world. Scientists should not present AI-generated content as observations collected in the real world.

Model creators should clearly identify, annotate, and maintain provenance about synthetic data used in their training procedures and monitor the issues, concerns, and behaviors arising from the reuse of computer-generated content in training future models.

4. **A focus on ethics and equity**
Scientists and *model creators* should take credible steps to ensure that their uses of AI produce scientifically sound and socially beneficial results while taking appropriate steps to mitigate the risk of harm. This includes advising scientists and the public on the handling of tradeoffs associated with making certain AI technologies available to the public, especially in light of potential risks stemming from inadvertent outcomes or malicious applications.

Scientists and *model creators* should adhere to ethical guidelines for AI use, particularly in terms of respect for clear attribution of observational versus AI-generated sources of data, intellectual property, privacy, disclosure, and consent, as well as the detection and mitigation of potential biases in the construction and use of AI systems. They should also continuously monitor other societal ramifications likely to arise as AI is further developed and deployed and update practices and rules that promote beneficial uses and mitigate the prospect of social harm.

Scientists, model creators, and *policymakers* should promote equity in the questions and needs that AI systems are used to address as well as equitable access to AI tools and educational opportunities. These efforts should empower a diverse community of scientific investigators to leverage AI systems effectively and to address the diverse needs of communities, including the needs of groups that are traditionally underserved or marginalized. In addition, methods for soliciting meaningful public participation in evaluating equity and fairness of AI technologies and uses should be studied and employed.

AI should not be used without careful human oversight in decisional steps of peer review processes or decisions around career advancement and funding allocations.

5. **Continuous monitoring, oversight, and public engagement**
Scientists, together with representatives from academia, industry, government, and civil society, should continuously monitor and evaluate the impact of AI on the scientific process, and with transparency, adapt strategies as necessary to maintain integrity. Because AI technologies are rapidly evolving, research communities must continue to examine and understand the powers, deficiencies, and influences of AI; work to anticipate and prevent harmful uses; and harness its potential to address critical societal challenges. AI scientists must at the same time work to improve the effectiveness of AI for the sciences, including addressing challenges with veracity, attribution, explanation, and transparency of training data and inference procedures. Efforts should be undertaken within and across sectors to pursue ongoing study of the status and dynamics of the use of AI in the sciences and pursue meaningful methods to solicit public participation and engagement as AI is developed, applied, and regulated. Results of this engagement and study should be broadly disseminated.

A New Strategic Council to Guide AI in Science

We call upon the scientific community to establish oversight structures capable of responding to the opportunities AI will afford science and to the unanticipated ways in which AI may undermine scientific integrity.

We propose that the National Academies of Sciences, Engineering, and Medicine establish a *Strategic Council on the Responsible Use of Artificial Intelligence in Science*.* The council

*Patterned after the existing Strategic Council for Research Excellence, Integrity, and Trust at the NASEM, this Strategic Council will also operate in a nimble, strategic, and responsive manner to address critical issues in a fast-moving area that impacts the conduct and trustworthiness of scientific research. The narrower focus on AI will allow this second Strategic Council to focus on impacts of AI and involve users, developers, and other stakeholders in the applications of AI to scientific advancement.

should coordinate with the scientific community and provide regularly updated guidance on the appropriate uses of AI, especially during this time of rapid change. The council should study, monitor, and address the evolving uses of AI in science; new ethical and societal concerns, including equity; and emerging threats to scientific norms. The council should share its insights across disciplines and develop and refine best practices.

More broadly, the scientific community should adhere to existing guidelines and regulations, while contributing to the ongoing development of public and private AI governance. Governance efforts must include engagement with the public about how AI is being used and should be used in the sciences.

With the advent of generative AI, all of us in the scientific community have a responsibility to be proactive in safeguarding the norms and values of science. That commitment—together with the five principles of human accountability and responsibility for the use of AI in science and the standing up of the council to provide ongoing guidance—will support the pursuit of trustworthy science for the benefit of all.

ACKNOWLEDGMENTS. David Baltimore, Caltech; George Q. Daley, Harvard Medical School; Harold Varmus, Weill Cornell Medical College; and Anne-Marie Mazza, National Academies of Sciences, Engineering, and Medicine made valuable contributions to this effort.

1. National Academies of Sciences, Engineering, and Medicine. *Fostering Integrity in Research* (The National Academies Press, Washington, DC, 2017), 10.17226/21896.
2. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science* (The National Academies Press, Washington, DC, 2019), 10.17226/25303.
3. National Academies of Sciences, Engineering, and Medicine. *Fostering Responsible Computing Research: Foundations and Practices* (The National Academies Press, Washington, DC, 2022), 10.17226/26507.
4. M. Aidinoff, D. Kaiser, "Novel technologies and the choices we make: Historical precedents for managing artificial intelligence" in *Issues in Science and Technology* (2024), 10.58875/BUXB2813.
5. U. Gasser, "Governing AI with intelligence" in *Issues in Science and Technology* (2024), 10.58875/AWJG1236.
6. M. L. Gray, "A human rights framework for AI research worthy of public trust" in *Issues in Science and Technology* (2024), 10.58875/ERUU8159.
7. A. J. London, "A justice-led approach to AI innovation" in *Issues in Science and Technology* (2024), 10.58875/KNRZ2697.
8. S. Parthasarathy, J. Katzman, "Bringing communities in, achieving AI for all" in *Issues in Science and Technology* (2024), 10.58875/SLRG2529.
9. E. Horvitz, T. M. Mitchell, "Progress in AI: History and Trajectory" in *Issues in Science and Technology* (in press), 10.58875/ZRFJ6746.
10. National Academies of Sciences, Engineering, and Medicine. *Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop* (The National Academies Press, Washington, DC, 2022), 10.17226/26532.