Comment

BING GUAN/BLOOMBERG VIA GETTY

Artificial-intelligence models require the vast computing power of supercomputers, such as this one at the University of California, San Diego.

Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research

Brooks Hanson, Shelley Stall, Joel Cutcher-Gershenfeld, Kristina Vrouwenvelder, Christopher Wirz, Yuhan (Douglas) Rao & Ge Peng

Artificial-intelligence tools are transforming data-driven science — better ethical standards and more robust data curation are needed to fuel the boom and prevent a bust.

Science is producing data in amounts so large as to be unfathomable. Advances in artificial intelligence (AI) are increasingly needed to make sense of all this information (see ref. 1 and *Nature Rev. Phys.* **4**, 353; 2022). For example, through training on copious quantities of data, machine-learning (ML) methods get better at finding patterns without being explicitly programmed to do so.

In our field of Earth, space and environmental sciences, technologies ranging from sensors to satellites are providing detailed views of the planet, its life and its history, at all scales. And AI tools are being applied ever more widely — for weather forecasting[2] and climate modelling[3], for managing energy and water[4], and for assessing damage during disasters to speed up aid responses and reconstruction efforts.

The rise of AI in the field is clear from tracking abstracts[5] at the annual conference of the American Geophysical Union (AGU) — which typically gathers some 25,000 Earth and space scientists from more than 100 countries. The number of abstracts that mention AI or ML has increased more than tenfold between 2015 and 2022: from less than 100 to around 1,200 (that is, from 0.4% to more than 6%; see 'Growing AI use in Earth and space science')[6].

Yet, despite its power, AI also comes

with risks. These include misapplication by researchers who are unfamiliar with the details, and the use of poorly trained models or badly designed input data sets, which deliver unreliable results and can even cause unintended harm. For example, if reports of weather events — such as tornadoes — are used to build a predictive tool, the training data are likely to be biased towards heavily populated regions, where more events are observed and reported. In turn, the model is likely to over-predict tornadoes in urban areas and under-predict them in rural areas, leading to unsuitable responses[7].

Data sets differ widely, yet the same questions arise in all fields: when, and to what extent, can researchers trust the outcomes of AI and mitigate harm? To explore such questions, the AGU, with the support of NASA, last year convened a community of researchers and ethicists (including us) at a series of workshops. The aim was to develop a set of principles and guidelines around the use of AI and ML tools in the Earth, space and environmental sciences, and to disseminate them (see 'Six principles to help build trust')[6].

Answers will evolve as AI develops, but the principles and guidelines will remain grounded in the basics of good science — how data are collected, treated and used. To guide the scientific community, here we make practical recommendations for embedding openness, transparency and curation in the research process, and thus helping to build trust in AI-derived findings.

Watch out for gaps and biases

It is crucial for researchers to fully understand the training and input data sets used in an AI-driven model. This includes any inherent biases — especially when the model's outputs serve as the basis of actions such as disaster responses or preparation, investments or health-care decisions. Data sets that are poorly thought out or insufficiently described increase the risk of 'garbage in, garbage out' studies and the propagation of biases, rendering outcomes meaningless or, even worse, dangerous.

For example, many environmental data have better coverage or fidelity in some regions or communities than in others. Areas that are often under cloud cover, such as tropical rainforests, or that have fewer *in situ* sensors or satellite coverage, such as the polar regions, will be less well represented. Similar disparities across regions and communities exist for health and social-science data.

The abundance and quality of data sets are known to be biased, often unintentionally, towards wealthier areas and populations and against vulnerable or marginalized communities, including those that have historically been discriminated against[7,8]. In health data, for instance, AI-based dermatology algorithms have been shown to diagnose skin lesions and rashes less accurately in Black people than in white people, because the models are trained on data predominantly collected from white populations[8].

Such problems can be exacerbated when data sources are combined — as is often required to provide actionable advice to the public, businesses and policymakers. Assessing the impact of air pollution[9] or urban heat[10] on the health of communities, for example, relies on environmental data as well as on economic, health or social-science data.

Unintended harmful outcomes can occur when confidential information is revealed, such as the location of protected resources or endangered species. Worryingly, the diversity of data sets now being used increases the risks of adversarial attacks that corrupt or degrade the data without researchers being aware[11]. AI and ML tools can be used maliciously, fraudulently or in error — all of which can be difficult to detect. Noise or interference can be added, inadvertently or on purpose, to public data sets made up of images or other content. This can alter a model's outputs and the conclusions that can be drawn. Furthermore, outcomes from one AI or ML model can serve as input for another, which multiplies their value but also multiplies the risks through error propagation.

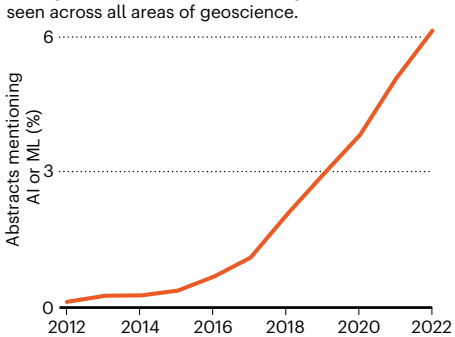
Our recommendations for data deposition (see ref. 6 and 'Six principles to help build trust') can help to reduce or mitigate these risks in individual studies. Institutions should also ensure that researchers are trained to assess data and models for spurious and inaccurate results, and to view their work through a lens of environmental justice, social inequity and implications for sovereign nations[12,13]. Institutional review boards should include expertise that enables them to oversee both AI models and their use in policy decisions.

Develop ways to explain how AI models work

When studies using classical models are published, researchers are usually expected to provide access to the underlying code, and any relevant specifications. Protocols for reporting limitations and assumptions for AI models are not yet well established, however. AI tools often lack explainability — that is, transparency and interpretability of their programs. It is often impossible to fully understand how a

GROWING AI USE IN EARTH AND SPACE SCIENCE

A rising proportion of abstracts for the annual meeting of the American Geophysical Union mention artificial intelligence (AI) or machine learning (ML) — a trend seen across all areas of geoscience.



SOURCE: REF. 5

Six principles to help build trust

Following these best practices will help to avert harm when using AI in research.

Researchers

1. Transparency. Clearly document and report participants, data sets, models, bias and uncertainties.
2. Intentionality. Ensure that the AI model and its implementations are explained, replicable and reusable.
3. Risk. Consider and manage the possible risks and biases that data sets and algorithms are susceptible to, and how they might affect the outcomes or have unintended consequences.
4. Participatory methods. Ensure inclusive research design, engage with communities at risk and include domain expertise.

Scholarly organizations (including research institutions, publishers, societies and funders)

5. Outreach, training, and leading practices. Provide for all roles and career stages.
6. Sustained effort. Implement, review and advance these guidelines.

More detailed recommendations are available in the community report[6] facilitated by the American Geophysical Union, and are organized into modules for ease of distribution, use in teaching and continued improvement.

EUROPEAN SPACE AGENCY/COPERNICUS SENTINEL DATA (2017)/SPL

AI tools are being used to assess environmental observations, such as this satellite image of agricultural land in Bolivia that was once a forest.

result was obtained, what its uncertainty is or why different models provide varying results[14]. Moreover, the inherent learning step in ML means that, even when the same algorithms are used with identical training data, different implementations might not replicate results exactly. They should, however, generate results that are analogous.

In publications, researchers should clearly document how they have implemented an AI model to allow others to evaluate results. Running comparisons across models and separating data sources into comparison groups are useful soundness checks. Further standards and guidance are urgently needed for explaining and evaluating how AI models work, so that an assessment comparable to statistical confidence levels can accompany outputs. This could be key to their further use.

Researchers and developers are working on such approaches, through techniques known as explainable AI (XAI) that aim to make the behaviour of AI systems more intelligible to users. In short-term weather forecasting, for example, AI tools can analyse huge volumes of remote-sensing observations that become available every few minutes, thus improving the forecasting of severe weather hazards. Clear explanations of how outputs were reached are crucial to enable humans to assess the validity and usefulness of the forecasts, and to decide whether to alert the public or use the output in other AI models to predict the likelihood and extent of fires or floods[2].

In Earth sciences, XAI attempts to quantify or visualize (for example, through heat maps) which input data featured more or less prominently in reaching the model's outputs in any given task. Researchers should examine these explanations and ensure that they are reasonable.

Forge partnerships and foster transparency

For researchers, transparency is crucial at each step: sharing data and code; considering further testing to enable some forms of replicability and reproducibility; addressing risks and biases in all approaches; and reporting uncertainties. These all necessitate an expanded description of methods, compared with the current way in which AI-enabled studies are reported.

> **"Recognized, quality-assured data sets are particularly needed for generating trust in AI."**

Research teams should include specialists in each type of data used, as well as members of communities who can be involved in providing data or who might be affected by research outcomes. One example is an AI-based project that combined Traditional Knowledge from Indigenous people in Canada with data collected using non-Indigenous approaches to identify areas that were best suited to aquaculture (see go.nature.com/46yqmdr).

Sustain support for data curation and stewardship

There is already a movement across scientific fields for study data, code and software to be reported following FAIR guidelines, meaning that they should be findable, accessible, interoperable and reusable. Increasingly, publishers are requiring that data and code be deposited appropriately and cited in the reference sections of primary research papers, following data-citation principles[15,16]. This is welcome, as are similar directives from

funding bodies, such as the 2022 'Nelson memo' to US government agencies (see go.nature.com/3qkqzes).

Recognized, quality-assured data sets are particularly needed for generating trust in AI and ML, including through the development of standard training and benchmarking data sets[17]. Errors made by AI or ML tools, along with remedies, should be made public and linked to the data sets and papers. Proper curation helps to make these actions possible.

Leading discipline-specific repositories for research data provide quality checks and the ability to correct or add information about data limitations and bias — including after deposition. Yet we have found that the current data requirements set by funders and journals have inadvertently incentivized researchers to adopt free, quick and easy solutions for preserving their data sets. Generalist repositories that instantly register the data set with a digital object identifier (DOI) and generate a supporting web page (landing page) are increasingly being used. Completely different types of data are too often gathered under the same DOI, which can cause issues in the metadata, make provenance hard to trace and hinder automated access.

This trend is evident from data for papers published in all journals of the AGU[5], which implemented deposition policies in 2019 and started enforcing them in 2020. Since then, most publication-related data have been deposited in two generalist repositories: Zenodo and figshare (See 'Rise in data archiving'). (Figshare is owned by Digital Science, which is part of Holtzbrinck, the majority shareholder in *Nature*'s publisher, Springer Nature.) Many institutions maintain their own generalist repositories, again often without discipline-specific, community-vetted curation practices.

This means that many of the deposited research data and metadata meet only two of the FAIR criteria: they are findable and accessible. Interoperability and reusability require sufficient information about data provenance, calibration, standardization, uncertainties and biases to allow data sets to be combined reliably — which is especially important for AI-based studies.

Disciplinary repositories, as well as a few generalist ones, provide this service — but it takes trained staff and time, usually several weeks at least. Data deposition must therefore be planned well before the potential acceptance of a paper by a journal.

More than 3,000 research repositories exist[18], although many are not actively accepting new data. The most valuable repositories are those that have long-term funding for storage and curation, and accept data globally, such as GenBank, the Protein Data Bank and the EarthScope Consortium (for seismological and geodetic data). Each is part of an international collaboration network. Some repositories are funded, but are restricted to data derived from the funder's (or country's) grants; others have short-term funding or require a deposition fee. This complex landscape, the various restrictions on deposition and the fact that not all disciplines have an appropriate, curated, field-specific repository all contribute to driving users towards generalist repositories, which compounds the risks with AI models.

Scholarly organizations such as professional societies, funding agencies, publishers and universities have the necessary leverage to promote progress. Publishers, for example, should implement checks and processes to ensure that AI and ML ethics principles are supported through the peer-review process and in publications. Ideally, common standards and expectations for authors, editors and reviewers should be adopted across publishers and be codified in existing ethical guidance (such as through the Council of Science Editors).

We also urge funders to require that researchers use suitable repositories as part of their data sharing and management plan. Institutions should support and partner with those, instead of expanding their own generalist repositories.

Sustained financial investments from funders, governments and institutions — that do not detract from research funds — are needed to keep suitable repositories running, and even just to comply with new mandates[16].

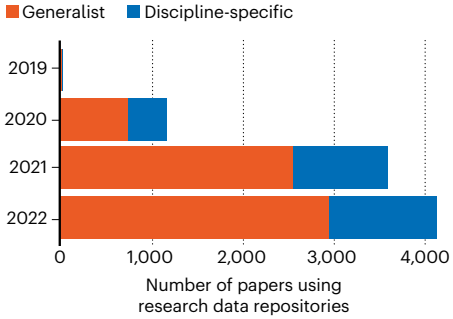
Look at long-term impact

The broader impacts of the use of AI and ML in science need to be tracked. Research that assesses workforce development, entrepreneurial innovation, real community engagement and the alignment of all the scholarly organizations involved is needed. Ethical aspects must remain at the forefront of these endeavours: AI and ML methods must reduce social disparities rather than exacerbate them; enhance trust in science rather than undercut it; and intentionally include key stakeholder voices, not leave them out.

AI tools, methods and data generation are advancing faster than institutional processes for ensuring quality science and accurate results. The scientific community must take urgent action, or risk wasting research funds and eroding trust in science as AI continues to develop.

RISE IN DATA ARCHIVING

Researchers are increasingly depositing data in repositories to widen access, but mostly in generalist rather than discipline-specific ones that offer curation. This can be seen in the top 15 repositories used in primary-research papers published in American Geophysical Union journals between 2019 and 2022.



SOURCE: REF. 5

The authors

Brooks Hanson is retired executive vice-president for science at the American Geophysical Union, Washington DC, USA. **Shelley Stall** is vice-president of Open Science Leadership at the American Geophysical Union, Washington DC, USA. **Joel Cutcher-Gershenfeld** is the Florence G. Heller Professor and director of the Social Impact MBA in the Heller School for Social Policy and Management at Brandeis University in Waltham, Massachusetts. **Kristina Vrouwenvelder** is program manager of Open Science at the American Geophysical Union, Washington DC. **Christopher Wirz** is a project scientist at the National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA. **Yuhan (Douglas) Rao** is a research scholar at the North Carolina Institute for Climate Studies, North Carolina State University, Asheville, North Carolina, USA. **Ge Peng** is a senior principal research scientist at the Earth System Science Center/NASA MSFC IMPACT, The University of Alabama in Huntsville, Huntsville, Alabama, USA.
e-mail: sstall@agu.org

1. National Academies of Sciences, Engineering, and Medicine. *Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop* (National Academies Press, 2022).
2. Hilburn, K. A., Ebert-Uphoff, I. & Miller, S. D. *J. Appl. Meteorol. Climatol.* **60**, 3–21 (2021).
3. Schneider, T. *et al. Nature Clim. Change* **13**, 887–889 (2023).
4. Rolnick, D. *et al. ACM Comput. Surv.* **55**, 42 (2022).
5. Vrouwenvelder, K. Preprint at Zenodo https://doi.org/10.5281/zenodo.8388025 (2023).
6. Stall, S. *et al.* Preprint at ESS Open Archive https://doi.org/10.22541/essoar.168132856.66485758/v1 (2023).
7. McGovern, A., Ebert-Uphoff, I., Gagne, D. J. & Bostrom, A. *Environ. Data Sci.* **1**, e6 (2022).
8. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D. & Tzovara, A. *Patterns* **2**, 100347 (2021).
9. Conibear, L. *et al. GeoHealth* **6**, e2021GH000570 (2022).
10. Shandas, V., Voelkel, J., Williams, J. & Hoffman, J. *Climate* **7**, 5 (2019).
11. Papernot, N., McDaniel, P. & Goodfellow, I. Preprint at https://arxiv.org/abs/1605.07277 (2016).
12. Pandya, R. *et al.* Preprint at ESS Open Archive https://doi.org/10.22541/essoar.167768122.22544063/v1 (2023).
13. Carroll, S. R. *et al. Data Sci. J.* **19**, 43 (2020).
14. McGovern, A. *et al. Bull. Am. Meteorol. Soc.* **100**, 2175–2199 (2019).
15. Data Citation Synthesis Group. *Joint Declaration of Data Citation Principles* (ed. Martone, M.) (FORCE11, 2014).
16. Stall, S. *et al. Nature* **570**, 27–29 (2019).
17. Thiyagalingam, J., Shankar, M., Fox, G. & Hey, T. *Nature Rev. Phys.* **4**, 413–420 (2022).
18. Pampel, H. *et al. Sci. Data* **10**, 571 (2023).

B.H. declares competing interests (see go.nature.com/3tpszhu).