PNAS PERSPECTIVE OPEN ACCESS

How should the advancement of large language models affect the practice of science?

Marcel Binz[a,b,1,2], Stephan Alaniz[b,c,d,2], Adina Roskies[e], Balazs Aczel[f], Carl T. Bergstrom[g], Colin Allen[e], Daniel Schad[h], Dirk Wulff[i,j], Jevin D. West[g], Qiong Zhang[k], Richard M. Shiffrin[l], Samuel J. Gershman[m], Vencislav Popov[n], Emily M. Bender[g,3], Marco Marelli[o,3], Matthew M. Botvinick[p,q,3], Zeynep Akata[b,c,d,4], and Eric Schulz[a,b,3,4]

Edited by David Kellen, Syracuse University, Syracuse, NY; received February 23, 2024; accepted October 21, 2024 by Editorial Board Member Elke U. Weber

Large language models (LLMs) are being increasingly incorporated into scientific workflows. However, we have yet to fully grasp the implications of this integration. How should the advancement of large language models affect the practice of science? For this opinion piece, we have invited four diverse groups of scientists to reflect on this query, sharing their perspectives and engaging in debate. Schulz et al. make the argument that working with LLMs is not fundamentally different from working with human collaborators, while Bender et al. argue that LLMs are often misused and overhyped, and that their limitations warrant a focus on more specialized, easily interpretable tools. Marelli et al. emphasize the importance of transparent attribution and responsible use of LLMs. Finally, Botvinick and Gershman advocate that humans should retain responsibility for determining the scientific roadmap. To facilitate the discussion, the four perspectives are complemented with a response from each group. By putting these different perspectives in conversation, we aim to bring attention to important considerations within the academic community regarding the adoption of LLMs and their impact on both current and future scientific practices.

large language models | AI | science

Language models are statistical models of human language that can be used to predict the next token (e.g., a word or character) for a given text sequence. Even though these models have been around for decades (1, 2), they have recently experienced an unprecedented renaissance: By training enormous neural networks with billions of parameters on datasets with trillions of tokens, researchers have observed the emergence of models whose abilities can go beyond mere text generation and conversational skills (3).

Modern large language models (LLMs) are, among other things, able to solve selected university-level math problems (4) by writing the code that calculates the solution, support language translation (5), or answer questions in a bar exam with high accuracy (6), out of the box and without additional training. Given the range of these capabilities, it seems possible that these systems will have an enormous impact on our society, leaving their mark on the labor market (7), the education system (8), and many other parts of our daily lives.

We—as scientists—may therefore wonder how will the advancement of LLMs affect the practice of science (9). Finding answers to this question is urgent as LLMs are already starting to permeate the academic landscape (10–17). For instance, in 2022, MetaAI released the first science-specific LLM (under the name Galactica) aimed to support researchers in the process of knowledge discovery (18). Even more recently, Terence Tao, a Fields Medal-winning mathematician, proclaimed (19) that "the 2023-level AI can already generate [. . .] promising leads to a working mathematician [. . .]. When integrated with tools such as formal proof verifiers, internet search, and symbolic math packages, I expect, say, 2026-level AI [. . .] will be a trustworthy co-author [. . .]."

Yet while there have been claims of immense potential for this technology for the advancement of science, there are also considerable concerns that need to be taken into account. For instance, the aforementioned Galactica model had to be taken offline after just three days because it was

Author affiliations: [a]Max Planck Institute for Biological Cybernetics, Tübingen, Baden-Württemberg 72076, Germany; [b]Helmholtz Center for Computational Health, Munich, Oberschleißheim, Bayern 85764, Germany; [c]Department of Computer Science, Technical University of Munich, München, Bayern 80333, Germany; [d]Munich Center for Machine Learning, München, Bayern 80333, Germany; [e]Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106; [f]Institute of Psychology, Eötvös Loránd University, Budapest 1053, Hungary; [g]Department of Linguistics, University of Washington, Seattle, WA 98195; [h]Psychology Department and Institute of Mind, Brain and Behavior, Health and Medical University, Potsdam, Brandenburg 14471, Germany; [i]Max-Planck-Institute for Human Development, Berlin 14195, Germany; [j]Center for Cognitive and Decision Science, University of Basel, Basel 4001, Switzerland; [k]Department of Psychology, Rutgers University, New Brunswick, NJ 08901; [l]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47408; [m]Department of Psychology, Harvard University, Cambridge, MA 02138; [n]Department of Psychology, University of Zurich, Zurich 8006, Switzerland; [o]Department of Psychology, University of Milano-Bicocca, Milano 20126, Italy; [p]Section/Unit, Google DeepMind, London N1C4AG, United Kingdom; and [q]Gatsby Computational Neuroscience Unit, University College London, London WC1E 6BT, United Kingdom

R.M.S. is an organizer of this Special Feature.

Author contributions: M.B. and S.A. project administration; Z.A. and E.S. project supervision; E.M.B., M.M., M.M.B., and E.S. Perspective leaders; A.R., B.A., C.T.B., J.D.W., Q.Z., E.M.B., M.M., M.M.B., and E.S. Perspectives/responses—original draft; all the authors Perspectives/responses—review and editing; M.B. and S.A. Introduction and conclusion—original draft; M.B., S.A., Z.A., and E.S. Introduction and conclusion—review and editing; and M.B., S.A., A.R., B.A., C.T.B., C.A., D.S., D.W., J.D.W., Q.Z., R.M.S., S.J.G., V.P., E.M.B., M.M., M.M.B., Z.A., and E.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. D.K. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: marcel.binz@helmholtz-munich.de.

[2]M.B. and S.A. contributed equally to this work.

[3]E.M.B., M.M., M.M.B., and E.S. contributed equally to this work.

[4]Z.A. and E.S. contributed equally to this work.

Published January 27, 2025.

heavily criticized by researchers for fabricating information, such as "fake papers (sometimes attributing them to real authors), and [...] wiki articles about the history of bears in space" (20). Furthermore, even though LLMs often achieve state-of-the-art performance on existing benchmarks, it remains debated whether this reflects genuine understanding, or whether they are merely acting like stochastic parrots (21). It has been, for instance, repeatedly demonstrated that even the most capable models at present fail at basic arithmetic problems such as multiplying two four-digit numbers (22) when directly asked for the answer as opposed to writing code. Flaws like these are especially concerning if we intend to utilize LLMs for research purposes and could endanger the integrity of science if we act carelessly.

The objective of the present article is to provide researchers with different opinions and a forum to voice and discuss their perspectives on if and how we should make use of LLMs in the context of science. To facilitate this discussion, the following section will first highlight a few applications where LLMs have the potential to positively impact science, followed by pointing out some of the issues that come with them.

Background: Applications of LLMs in Science

LLMs find their most obvious use case as a supporting tool for scientific writing. For example, they could provide starting points for letters of recommendation or evaluation, search for and summarize relevant research, and review or edit journal submissions. When used as proofreaders of manuscript drafts, they can aid in rectifying grammatical errors, improving the writing style, and ensuring adherence to editorial guidelines. Beyond scientific writing, LLMs could prove valuable for data acquisition and analysis in domains that were traditionally reliant on manual human work (23, 24). Researchers have even suggested using LLMs as potential substitutes for human participants, as proxies (25) or for pilot studies (26). In such settings, it has been argued that LLMs could augment human data or help to gauge the effects of intended experimental manipulations before conducting the study in the wild, thereby saving time and money (27). In computational fields, LLMs could speed up prototyping by proposing code (28), while a human-in-the-loop would guide these processes, correct LLM-generated errors, and ultimately decide which ideas warrant further pursuit. Moreover, researchers might experiment with employing LLMs at certain stages of research with progressively reduced supervision (29), potentially leading to increased automation in some aspects of scientific exploration and discovery.

While the potential influence of LLMs on the practice of science is immense, there are pressing issues that come with the use of LLMs in the context of science. When an LLM helps us to write text, who ensures that its output is not subject to plagiarism issues (30)? LLMs learn from web-sourced text data, acquiring inherent biases (31–33) and—in some cases—replicate excerpts from their training data (34). The New York Times, for instance, recently filed a lawsuit against Microsoft and OpenAI for unlawful use of its articles to create LLMs, thereby highlighting legal issues surrounding such practices (35). When an LLM is used for data analysis, what happens when it makes up or changes data? The content generated by LLMs can contain errors or fabricated information, presenting a potential threat to the integrity of scientific publishing (13). When an LLM suggests an idea, who gets credit for it? The general consensus within the scientific community seems to indicate that LLMs are not eligible for (co-)authorship (36) as they cannot be held accountable for upholding scientific precision and integrity. Leading AI conferences such as ICML* and ACL†—as well as journals such as Science,‡ Nature§ and PNAS¶—have already adopted policies to limit the involvement of LLMs. However, it remains an open question how strong these regulations should be and if and how the usage of LLMs should be acknowledged.

These—and many other—issues raise the questions: How *should* the advancement of LLMs affect the practice of science? Do LLMs actually improve our scientific output or are they rather hindering good scientific practice? To what extent should they be used given the ethical and legal issues that come with them? We believe these to be highly non-trivial questions without an obvious answer and have therefore invited four groups of researchers to provide their perspectives on them. These perspectives were selected to cover a broad spectrum of opinions in order to spark a constructive discussion. Each of the perspectives is accompanied by a response from each group. We conclude this article with a short general discussion in which we attempt to identify common themes.

Perspective—LLMs: More Like a Human Collaborator than a Software Tool

Contributors. Eric Schulz, Daniel Schad, Marcel Binz, Stephan Alaniz, Ven Popov, and Zeynep Akata.

Most researchers in our labs already frequently employ LLMs in their everyday work. They use them, among other things, to finetune and revise their drafts, as a supporting tool for programming, to suggest formulations for research items such as questionnaires or experimental instructions, and to summarize research papers. We have observed a significant increase in quality in all of these areas after the widespread adoption of these models. While our personal experience may be biased, there are several studies supporting the idea that LLMs can facilitate writing (37), coding (38), and knowledge extraction (39). In the future, we expect these models to be even more deeply integrated into the scientific process, taking on roles similar to a collaborator with whom one can develop and discuss ideas.

Indeed, we believe that working with LLMs will not be fundamentally different from working with other collaborators, such as research assistants or doctoral students. LLMs are not perfect and have limitations and biases that could affect their performance and output. However, humans are also subject to some of the same flaws, such as errors, plagiarism, fabrication, or discrimination. If we take this perspective, it seems appropriate to view current LLMs less

*https://icml.cc/Conferences/2023/llm-policy.

†https://2023.aclweb.org/blog/ACL-2023-policy/.

‡https://www.science.org/content/page/science-journals-editorial-policies.

§https://www.nature.com/nature-portfolio/editorial-policies/ai.

¶https://www.pnas.org/author-center/editorial-and-journal-policies#authorship-and-contributions.

as traditional software tools and more as knowledgeable research assistants: They can do phenomenal work but we need to be aware that they can make mistakes.

Protecting the past. It is our chief responsibility to ensure the quality and integrity of our work. There are already rules and norms about scientific practice in place to ensure this, and many of them also apply to LLMs. For instance, we should always check the accuracy and validity of the information and data we obtain, no matter the source, as well as correctly cite the sources and methods we use. That means that we should not blindly trust or rely on LLMs, but rather use them as a complement to our own expertise and judgment. Furthermore, our work can only be criticized appropriately if all information about its methodology is transparently communicated. We should therefore acknowledge the contributions of LLMs to our research, just as we would do for any other tool. Ultimately, it is—and will remain—the authors' responsibility to ensure that the appropriate scientific standards are followed, regardless of whether we use LLMs or not.

Ensuring that our research is reproducible is one of the cornerstones of modern science. However, as many LLMs are proprietary, working with them poses a threat to this ideal. Nobody guarantees that OpenAI, Google, or other providers will not make changes to their models (in the worst case, without informing the user). In fact, this happened to us during the revision process of one of our papers, where, at some point, we could not reproduce our initial results, likely due to changes on the provider side. Likewise, this has been observed by Yax et al. (40) who tested the reasoning capabilities of LLMs, and found that the results of proprietary LLMs, i.e., ChatGPT and GPT4, could not be replicated three months after the initial experiments. Their analyses also found that, surprisingly, the scores for some of the tests significantly decreased, exemplifying the reliability issues with proprietary LLMs. Inconsistencies like this can be an issue when analyzing the behavior of such models. How should we deal with cases like this? We believe that the obvious solution to this problem is to rely on open-source models (41, 42) where one has full control over all aspects of the model, i.e., they can be run locally and are clearly identified by their release version to ensure reproducibility. Following a recent call for action to the European Parliament (43), we therefore strongly advocate for the development of such models, such that they can become the primary tool for scientific inquiry as they are rapidly catching up with state-of-the-art proprietary models (42).

Welcoming the future. Paper reviewing is another area where LLMs could improve our scientific pipeline. In a recent study, Liang and colleagues (44) demonstrated this potential by systematically evaluating the quality of LLM-generated reviews. They invited researchers to submit their own papers and asked them—after having received an LLM-generated review—to judge its helpfulness relative to reviews they had received from human researchers. Their result indicates that "more than half (57.4%) of the users found GPT-4 generated feedback helpful or very helpful and 82.4% found it more beneficial than feedback from at least some human reviewers." Not only does this result allow scientists—especially early career researchers—to receive high-quality, instantaneous feedback (similar to that one could get from a critical colleague with an unlimited amount of time) but it also has implications for the peer review process. Yet, the use of LLMs in the peer review process also presents one major legal obstacle: Manuscripts under review are typically confidential, and hence should not be entered into proprietary LLMs. To prevent such breaches of confidentiality, the NIH and other institutions have rules in place that prohibit the use of LLMs for peer review.[#] Locally hosted, open-source models are again a solution to this issue, as they provide control about which information is shared with external sources and which is not.

We also would like to point out that LLMs are a moving target, constantly evolving and becoming more capable and autonomous. This may raise new challenges and questions for the scientific community in the future, such as how to evaluate, interpret, and communicate the results generated by LLMs, or how to ensure their transparency and accountability. We welcome these challenges as an opportunity to advance our understanding and methods of science. We also encourage researchers to collaborate with each other and with LLM developers to address these issues and ensure that LLMs improve at frequently criticized skills such as providing truthful sources or acknowledging ignorance.

Conclusion. In conclusion, LLMs are a valuable asset for science and should be embraced rather than feared or restricted. It becomes apparent that they are not infallible machines once we start thinking about them as knowledgeable research assistants instead of traditional software tools. Furthermore, since rules for good scientific practice are already in place, and since it is the authors' obligation to take responsibility for adhering to these rules, there is no need for novel rules with the use of LLMs. We believe that strengthening the development of open-source alternatives should be one of our top priorities, as they "offer enhanced security, explainability, and robustness due to their transparency and the vast community oversight" (43). Finally, being conscious about the current limitations of LLMs and embracing them, will allow us to grow with the technology as LLM research finds remedies and develops complementary tools. We hope that by adopting this liberal perspective, we can foster a positive and fruitful relationship between humans and LLMs in science. As an illustration of the way LLMs can be used productively as an "assistant," the first draft of our perspective was written by an LLM (GPT-4; accessed on September 22, 2023) based on our meeting notes.

Perspective—Science Is a Social Process That Cannot Be Autocompleted

Contributors. Emily M. Bender, Carl T. Bergstrom, and Jevin D. West.

When deciding whether to use an LLM, it is important to recognize that LLMs are simply models of word form distributions extracted from text—not models of the *information* that people might get from reading that text (45). Despite the attendant excitement, these systems aren't any closer to replicating human intelligence than the systems Dreyfus

[#] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html.

critiqued (46). Originally, language models were used to rank or classify text. In automatic transcription, for example, an acoustic model provides a set of possibilities and the language model helps determine the most likely next word (47). Today, however, LLMs are vaunted for their ability to extrude synthetic text by repeatedly selecting a next likely token.

Trained on sufficiently large datasets and with sufficiently well-tuned architectures and training processes, LLMs appear to produce coherent text on just about any topic, including scientific ones. Moreover, we are vulnerable to mistaking it for useful or informative text because our linguistic processing capabilities are instinctual and reflexive (48). In other words, we are ill-positioned to effectively evaluate LLM output because we can't help but make sense of it.

Proponents argue that LLMs are useful in three domains: 1) navigating science, by searching and synthesizing published literature, 2) doing science, in the sense of designing or conducting experiments or generating data, and 3) communicating science, by drafting text for publication. While certain machine approaches may be useful in each, we believe that LLMs are unlikely to outperform alternative technologies. Furthermore, we have serious concerns about the downstream potential for harms to science if their use is widely embraced.

Navigating science. Natural language processing (NLP) has proven useful in sorting through an ever-growing body of scientific literature. Information retrieval and extraction techniques, as implemented in academic search engines (e.g. ref. 49), have helped researchers discover relevant prior work. Will LLMs supplant other NLP approaches? We doubt it. The inappropriateness of LLMs as text generators and synthesis machines was highlighted in Meta's Galactica debacle. That system—taken off-line after three days in response to intense criticism for its abysmal performance—had been trained on scientific text and promoted as a tool to "summarize academic papers, solve math problems, generate Wiki articles, write scientific code, annotate molecules and proteins, and more" (20). But training an LLM on scientific papers doesn't guarantee that it will output scientifically accurate information. As Meta discovered, the use of LLMs yields text ungrounded in any communicative intent or accountability for accuracy.

One might hope that LLMs could at least be used to summarize a set of papers. Extractive summarization systems (50) already do this; will LLMs perform better? Will people tend to overrely on system output rather than using it as a starting point? What are the costs of false negatives, i.e., important points not included in the generated summary? How will errors generated by LLMs, which then become training data for future LLMs, get amplified?

Doing science. LLMs are just one of many technologies dubbed "AI," but their surprising capacity to perform what amounts to a fancy parlor trick has drawn outsized attention. That's a mistake. LLMs may be adequate for specific linguistic tasks such as grammar checking, automatic transcription, and machine translation (including code generation), but we anticipate that they will not prove as effective as other tools for most tasks involved in hybrid human-machine science. Even where they do appear to be moderately effective, they are known to be brittle to input variation (51). We envision the future of machine-aided science not as a massive, one-size-fits-all, universal application of LLMs, but rather an ensemble of bespoke and often lightweight models that have been designed explicitly to solve the specific tasks at hand—and, crucially, evaluated in terms of those specific tasks. Such approaches also have a major advantage where interpretability is concerned. If researchers want to understand output variation, let alone find ways to fine-tune the architecture to generate better results, they need to steer away from technologies as opaque as LLMs. But instead, the ongoing hype around LLMs is drawing funding and brainpower away from more promising, targeted approaches.

Not only are LLMs being explored as aides to researchers; numerous proposals suggest that they can stand in for test subjects (52), survey participants (53), or data annotators (54). Such arguments derive from a failure to understand that LLMs output sequences of linguistic tokens, not concepts, meanings, or communicative intent (45). If we are looking to study the opinions or behavior of human beings, we need to work with actual people.

Communicating science. By design LLMs generate form without substance. The synthetic text that systems output constitutes neither ideas, nor data—and it certainly is not a reliable information source. This notion of generating statements that no one intended is anathema to the spirit of scientific inquiry. Automatically generating something that looks like a manuscript is very different from the iterative process of actually writing a manuscript. Yet the output can be difficult to distinguish, particularly in a cursory read or by inexpert readers. Some proponents argue that LLMs can relieve scientists of the drudgery of writing papers and free them up to get on with the serious business of "doing science" (55). This false dichotomy between communication and investigation reflects a fundamental misunderstanding of the nature of science (56) that devalues the communicative aspects of science and ignores the role of writing in the process of formulating, organizing, and refining ideas.

Downstream, LLMs threaten the notion of scientific expertise, shift incentive structures (57), and undermine trust in the literature. Notions of systematic review are undercut by the randomness inherent in LLM output. And most importantly, when someone uses an LLM to generate a literature review, the claims generated are not directly derived from the manuscripts cited. Rather, the machine creates textual claims, and then predicts the citations that might be associated with similar text. Obviously, this practice violates all norms of scholarly citation. At best, LLMs gesticulate toward the shoulders of giants.

Driven by quantitative metrics and the strong incentive to publish, researchers may opt to trade off quality for speed by letting LLMs do much of their writing. Widespread use of one or a few LLMs could undercut epistemic diversity in science. When asked to provide a hypothesis, experiment, or mode of explanation, LLMs may repeatedly offer similar solutions, instead of leveraging the parallel creativity of an entire science community.

Worse still, opportunistic or malicious actors could use LLMs to generate nonsense at scale with minimal cost. (This is not an argument against using LLMs appropriately, but we need to be prepared for such behavior.) Lazy authors could boost their publication counts by shotgunning machine-

generated papers to low-quality journals. Predatory publishers could feign peer review using LLM output. Bad actors could overwhelm the manuscript submission system of a target journal (or even a target field) with a massive volume of fake papers. Or an investigator's work could be targeted with a deluge of spurious machine-generated critiques on postpublication peer review platforms such as Pubpeer.

Finally, LLMs may cause considerable collateral damage to science education. For example, as LLMs slash the cost of generating seemingly authoritative text, the web will be flooded with low-quality, error-ridden tutorials designed to capture advertising revenue. At present, search engines' ability to discriminate is more or less the only line of defense. That's worrisome.

Conclusion. In conclusion, LLMs are often mischaracterized, misused, and overhyped, yet they will certainly impact the way we do science, from search to experimental design to writing. The norms that we establish now around their use will determine the consequences far into the future. We should proceed with caution—and evaluate at every step.

Perspective—LLMs in Scientific Practice: A Matter of Principles, Not Just Regulations

Contributors. Marco Marelli, Adina Roskies, Balazs Aczel, Colin Allen, Dirk Wulff, and Qiong Zhang.

A moderate perspective on the potential impact of LLMs on scientific practice holds that, while it is important to be mindful of the dangers, their application seems largely beneficial insofar as they offer much needed support in day-to-day research activity and may alleviate major obstacles to scientific advancement. This is evident when LLMs are applied as editing tools: They provide a writing aid that leaves researchers with more time for brainstorming ideas and analysis, may help mitigate disparities between different scientific communities, and remedy some of the disadvantages for researchers who are not native speakers of English (58). Also, LLMs can access a broader range of literature than any individual researcher could, potentially offering valuable support for literature analysis and hypothesis generation (59), with a reach that, biases aside, goes beyond one's research specialization.

However, although any new technology may be harmful or helpful, some technologies afford opportunities for harm or help more than others. The reliance on LLMs has disruptive potential that is ever more evident, and such disruption must be kept at bay if the goal is to prevent "evil drifts." One perspective might hold that strict regulation is required, limiting the application of such systems at different points of the research endeavor, but regulation carries with it many costs that might be best avoided if kept moderate. A preferable approach may be to adopt clear principles guiding the way this technology should be used, principles that cannot just focus on efficiency and overall utility. Such principles include transparency, accountability, and fairness.

A matter of transparency. In science, transparency is of indispensable value. When used as writing tools, researchers must acknowledge the reliance on LLMs so that readers are on notice that the text is (at least partially) AI-generated. Authors should make explicit which LLMs were applied and how, as part of the method sections or in a separate dedicated statement. This could be achieved by relying on already existing solutions; for example, the CrediT taxonomy[‖] could also be used to code the nature of AI contribution, even if AI is not to be recognized as a coauthor. Ideally, in the spirit of open science, when possible and feasible, authors shall publicly release their prompts along with the corresponding LLM responses as supplementary materials, and reference such archives in the manuscript. One might question the value of doing so given that authors are already required to take responsibility for the content of their articles. But given the opacity of the relationship between an LLM's training data and its outputs, no author can fully verify that LLM-generated text is properly sourced. Readers deserve to be warned. One might further question our suggestion on the grounds that it could stigmatize authors who are not native speakers. However, indicating how the LLM was used, whether merely to clean up text supplied by authors or generate text from other kinds of prompts, would mitigate this concern and provide readers with important source information.

Importantly, transparency does not only pertain to the way we exploit LLMs, but to the systems themselves. LLMs are not, strictly speaking, anything new. Models that are analogous to current LLMs in structure, spirit, and basic mechanisms have been part of the scientific debate for decades (60). However, such older models were unambiguous about their architecture and training, if not openly released. Current LLMs are often not held to the same scientific standards as their ancestors, being widely applied even when their inner workings and training data remain undisclosed. This makes it difficult to estimate the actual performance of such models [and, importantly, the possibility of data contamination (61)]. As a scientific community valuing greater transparency, we should favor systems taking steps in that direction (62).

A matter of accountability. It must be acknowledged that LLMs are instruments of human agency, and researchers should be held accountable for any scientific product they present to the community, irrespective of the extent to which this was obtained through the application of automatic systems. The Association for the Advancement in AI has released clear guidelines in this respect: "Attribution of authorship carries with it accountability for the work, which cannot be effectively applied to AI systems ... Ultimately, all authors are responsible for the entire content of their papers, including text, figures, references, and appendices." For example, LLMs are known to "hallucinate" and produce factually incorrect responses (63). They fabricate bibliographic citations, omit important references when summarizing literature, and potentially plagiarize text written by another researcher. Even if this scenario is rapidly changing, and factuality is a central issue in current developments, the burden to verify that LLM-produced texts are accurate and that LLM-proofread texts are consistent with the original message remains with the individual authors. Similarly, LLMs' performance in logic and deductive tasks is often poor (64, 65), so using them for analysis may lead to false conclusions. The onus is on the user to ensure that what LLMs produce is worth pursuing. Researchers must hence have strategies for the assessment of AI-related content;

[‖] https://credit.niso.org/.

a good practice would be to have clear quality criteria and verification methods defined before using LLMs, as it is already standard practice in the programming industry. Scientists should not underestimate the time and effort that such vetting will take and should weigh the efficiency of LLM application against these costs. The time saved in text generation might be offset by the time required to verify the text generated.

A matter of fairness. The diffusion of AI systems in general and LLMs in particular have the potential to deeply affect us at a societal level. Science, as any human endeavor, is not immune to this. As a community, we must make all possible efforts to guarantee that reliance on LLMs does not violate basic fairness principles. Indeed, current language models reflect mostly WEIRD (Western Educated Industrialized Rich Democratic) populations and cannot easily be prompted to represent non-WEIRD communities (66, 67). This leads to biases in writing and annotation, potentially reinforcing distortions in citations and marginalization of already marginalized scientists. It also could lead to biases in communicating and interpreting results relevant to social, ethical, and political values that impact individual and public decision-making (68). Moreover, the wide application of LLMs may have negative consequences in terms of equitable research; in fact, such systems are also more accessible to WEIRD populations and, even within WEIRD countries, there could be a lot disparities in the ability to access the best versions of such technology, which are typically behind a paywall. These systematic patterns must be recognized and taken into account, to avoid unprincipled biases affecting the direction of research and possibly the relative success of careers. More generally, being aware of such biases (by, for example, adapting LLM prompts) can help alleviate their impact on society as a whole.

Conclusion. The impact that LLMs are having on scientific practice cannot be understated. Given the current trend, at the time you are reading these words, such impact will likely be much larger than it is as we write this piece. How the advancement of LLMs will influence the practice of science in the future cannot be entirely predicted; countering such a revolution with strict, preconceived norms is a losing battle. Rather, establishing principles and shared values in the scientific community constitute the ideal foundation for managing these rapidly changing technologies. A healthy skepticism is a pillar of any scientific enterprise. We need to train students and each other to build upon these principles in order to become appropriately skeptical toward LLMs and their outputs.

Perspective—AI Can Help, But Science Is for People

Contributors. Matthew M. Botvinick and Samuel J. Gershman.

Like many forms of technology, AI can substitute for human work. With the advancement of LLMs, the relevant kinds of work begin to overlap with high-level human cognitive work, including the activities involved in science (15). As LLMs improve, their ability to substitute for human scientific work will be a major boon. However, we argue here that two core aspects of scientific work should be reserved to human scientists.

AI and scientific work. Over time, the work involved in scientific research has become progressively more onerous, sometimes now bordering on the intractable. Assimilating current knowledge has become more difficult in the face of increasingly voluminous literatures. Generating new questions, hypotheses, and experimental tests has become more challenging, as the search problem entailed by each has become more complex. Drawing conclusions from experimental results has become harder as the size and complexity of datasets has exploded. And communicating and debating scientific conclusions has become more challenging for reasons including an overtaxing of peer review systems (69). Given the increasing costs of scientific work on these fronts, it's no surprise that progress across multiple scientific fields appears to have slowed (70).

In the long run, AI may help us cope with the increasing demands of scientific work. Through the kinds of application detailed in the introductory essay above, AI may help us scale up, by making each step in the research cycle cheaper. In some cases, AI may eventually perform some forms of scientific work better than human scientists, including the work of generating new hypotheses (71). Even in present-day forms, AI may be useful on some fronts, as reviewed in the introduction. Of course, as widely discussed, current systems are too unreliable to deploy without caution and oversight (see accompanying commentaries), and only time will tell how feasible it may be to overcome current limitations.

However, in addition to addressing present-day shortcomings, it's equally important to look into the future and consider what kind of AI tools we actually want for science in the long run. Given that AI can be applied to all phases of scientific work, one aim might be to build a full-fledged AI scientist, one that can do everything a human scientist now does: a full-spectrum replacement for human scientists. To us, this prospect is deeply unappealing. Why? Because there are particular aspects of science that we simply would not want to delegate to AI, even in a scenario where technical limitations presented no barrier. In particular, there are two core aspects of science that should be left to people. As we now explain, one of these is normative and the other epistemic.

The normative aspect of science. Any scientific discipline must continually ask, What problems shall we work on? How this question gets answered, both within individual labs and across whole research communities, is a complex affair, but it centers on judgments concerning the 'interest' and 'significance' of candidate problems, as well as their 'timeliness,' including their amenability to study under prevailing material and ethical constraints. Such judgments are informed by hard data; we obviously cannot reduce them to purely social constructions. However, at the same time, judgments of interestingness, significance, and timeliness are inherently tied to culturally and historically grounded sensibilities and mores. This is not a corruption or impurity in scientific thought and procedure. Cultural sensibilities and patterns of thought are fundamental to scientific prioritization.

This point will be especially salient to students of the history of science, because the sensibilities and mores that inform science evolve over time. Just as scientific theory changes over the years, so do the ethical com-

mitments and intellectual priorities that underlie science. This is evident in the fact that we no longer approach homosexuality as a disorder (72), or study genetics through the lens of eugenics. It shows in growing restrictions on animal experimentation (73). And it shows in the attention that Western climatologists now pay to regions historically neglected (74).

We argue that the normative aspect of science should not be ceded to AI systems, no matter how capable those systems become. People should stay in the driver's seat, determining the direction of travel for science. Certainly, AI systems may be helpful partners in deliberation, especially as techniques for AI value alignment improve (75). However, aligning a system to currently prevailing human views is different from allowing that system to govern the evolution of human views. In science, the ultimate driving force in that evolution should remain human. We are the moral agents in the room, and we shouldn't forget it.

The epistemic aspect of science. A central goal of basic science is understanding the natural world. If we are going to do science with AI tools, the question arises: "whose" understanding matters? Would it be satisfactory to have AI systems that successfully model aspects of nature—as reflected, for example, in accurate predictions—but which do not directly advance human knowledge concerning the underlying principles or mechanisms? From an engineering standpoint that might be fine. However, if it's basic science we're talking about, we shouldn't let go of the core objective, which is not just practical but epistemic. We cannot cede understanding to artificial systems. We should insist on human understanding remaining a core goal of science.

Of course, it may be that because of limitations on human cognition, AI systems may someday be able to represent some aspects of nature that we cannot, just as existing AI systems master aspects of complex board games that elude even highly skilled human players (76). Even in these cases, however, we should strive to extract as much human insight from AI systems as possible (77). We shouldn't lose track of what basic science is for. This doesn't preclude the use of AI advances in prediction for aiding human insight; prediction systems like AlphaFold are currently being used to advance basic science. Our point is that the intrinsically human objective of basic science cannot be entirely subsumed by predictive technology.

Conclusion. AI promises to deliver great value in science, just as in many other domains. We believe its potential should be embraced. However, at the same time that we strive to break through the current limitations of AI to access its benefits, we should also think through our long-term goals in developing this technology. In the end, the two areas of science we've proposed to protect—one normative, the other epistemic—are two reflections of a more general bound on AI's proper domain. We might call this the subjective limit. Unlike AI systems, people have a 'point of view,' which cannot be automated because it's inherently subjective (78). This point of view includes knowledge that is meaningful to us (the epistemic view) and values that are meaningful to us (the normative view). Machines might have their own knowledge or values, and these might be aligned with ours, but the alignment problem is fundamentally yoked to our subjective views. This principle applies in science, as in all human-centered activities.

Response by Eric Schulz, Daniel Schad, Marcel Binz, Stephan Alaniz, Ven Popov, and Zeynep Akata

We have argued that one should think of working with LLMs less as using a traditional software tool and more as working with a human collaborator and that this perspective allows us to better understand their shortcomings. This view actually resonates with many of the points raised in the other perspectives. For example, Marelli et al. write that "we should not blindly trust or rely on LLMs, but rather use them as a complement to our own expertise and judgment," and Bender et al. argue that collaboration in science means iterating over outputs many times. Like working with a human collaborator, working with LLMs is an iterative process in which we constantly check for facts and logical consistency, revise arguments, and identify new connections. This process takes time and is more than just booting up an LLM and copy-pasting its outputs; as nicely put by Marelli et al., we "should not underestimate the time and effort that such vetting will take, and should weigh the efficiency of LLM application against these costs."

However, we would also like to stress that the notion that "LLMs are simply models of word form distributions extracted from text" oversimplifies both their capabilities and the additional engineering effort involved in modern LLMs. If one takes steps like reinforcement learning from human feedback (79) or instruction tuning (80) out of the equation, the outputs produced by such models are rather uninspiring (anyone who has ever worked with a plain LLM can attest to this). Yet, with those ingredients included, LLMs do not just mimic language patterns; they can also synthesize concepts, critically evaluate their own outputs, and assist in problem-solving by processing vast amounts of data.

Bender et al. argue that the future of machine-aided science will not be "a massive, one-size-fits-all, universal application of LLMs, but rather an ensemble of bespoke and often lightweight models that have been designed explicitly to solve the specific tasks at hand [...]." We believe LLMs are widely adopted precisely because they are a universal tool to accomplish many tasks. Not only does that remove the need to build specialized tools for each application, but it also eradicates the time it takes to learn and adopt them. Like human collaborators, who bring a diverse range of skills to a project, LLMs offer a breadth of knowledge that can be tailored to specific needs, e.g., as shown with the finetuning of coding LLMs (28). There are—of course—applications that benefit from purposefully designed tools, but we believe that the percentage of such applications is modest once we take the time required to develop and adopt such tools into account.

Finally, there is the question of how much autonomy we want to transfer to LLMs or other AI systems. Botvinick and Gershman advocated that people should retain control over certain aspects of the scientific pipeline, such as deciding which topics to work on. We do not think that such a constraint is necessary. For example, if in the future, we can

employ an LLM (or any other AI system) to work on a topic that it deems interesting, and this LLM has proven itself to select topics in a very fruitful and productive manner, should we refrain from it? We do not think so as long as ethical and legal guidelines are followed. Deciding on scientific topics is hard, and it is often not a priori known which research directions will be fruitful. Therefore, we should take any help we can get. Human researchers and AI systems bring complementary strength to the table, and acknowledging this collaborative spirit enables us to leverage the best out of both worlds.

Response by Emily M. Bender, Carl T. Bergstrom, and Jevin D. West

Science is a social process. It cannot be autocompleted. Its agents—real scientists—are as much the product of this process as the results recorded in papers.

LLM optimists envision a new world, where machines write, review, and even do much of the science. Even the less extreme narrative wherein LLMs simply aid researchers suffers from a misplaced and almost Taylorist (81) optimism regarding production efficiency. Science is not a factory, churning out widgets or statistical analyses wrapped in text. For a factory, producing one more car per day is progress. For science, the goals are to understand our world—not to produce more artifacts that look like scientific papers. If science were a paper factory, we too would indulge in LLM euphoria and might even claim a significant resulting improvement in quality coming out of our labs. But we cannot equate papers and progress. Papers are but messages that we send one another to coordinate our collective quest for scientific understanding.

We don't, however, believe that any new mandates are required prohibiting the use of LLMs. All ill-advised use cases are already contrary to the norms of science: Using LLMs as stand-ins for human subjects (82) amounts to fabricating data; using LLMs to write first drafts runs afoul of prohibitions against plagiarism, as it is impossible to discern the source of any string produced by an LLM; treating LLMs as co-authors contravenes norms around authorship, since LLMs are not the sort of thing that can be accountable for paper contents; using LLMs to produce peer reviews is tantamount to abrogating our responsibility to deeply evaluate the methods, reasoning, and conclusions of our peers' work.

When contemplating how LLMs will affect science, we should not underestimate the temptation to use them under deadline pressure or in response to publish-or-perish threats to job security. Nor should we underestimate the time needed to fact-check all LLM output—not only for the inevitable and frequent errors but also to assess whether citations are accurate. We note that there are no published user studies that quantify just how much effort this checking process is, nor how accurately researchers can carry it out, especially while working under pressure. Norms of plagiarism and the weight of reputation will hopefully counterbalance the unfettered use of this new technology.

To reason appropriately about when LLMs are suitable within science, it is critical to avoid anthropomorphizing them. These models aren't research assistants. They are tools. They don't make mistakes like junior (or senior!) researchers do: People can take responsibility for, and learn from, mistakes. Tools produce errors; thus people using the tools have a responsibility to understand their affordances and use them with care.

Similarly, understanding LLMs as tools positions us to ask: Is this the best tool for this task? Often, we expect, LLMs are not. Even setting aside the closed proprietary models, their attendant failures of transparency, and the stochastic nature of LLM output, we expect that bespoke models designed for specific tasks will be more efficient, performant, interpretable, and easier to fix when not functioning well.

Ultimately, science is a conversation and the interlocutors are the scientists. Synthetic text-extruding machines, designed only to produce plausible-sounding prose, are not fit participants in that conversation and should not be treated as such.

Response by Marco Marelli, Adina Roskies, Balazs Aczel, Colin Allen, Dirk Wulff, and Qiong Zhang

We think that LLMs can be profitably incorporated into scientific practice (in line with Schulz et al.), but we also recognize that there are causes for reservation (in line with Bender et al.).

We disagree with the view that LLMs should be considered collaborators or research assistants (Schulz et al.). One can instruct students or research assistants, address their mistakes, and anticipate that they will learn from them. One may also question their reasons or reasoning and get answers and expect accountability. Finally, one may also get insight into their values and their motivations, and trust or distrust them accordingly. LLMs are not introspective, lack metacognition, and have no values, at least not in the way humans do. Indeed, our inability to understand why they make the errors they do or when they will make them impairs our ability to understand their limits, especially on the edges of knowledge, where their training corpus is arguably less robust (83, 84). Moreover, although LLMs move from the same foundations of previous language models (Schulz et al.), they are significantly more opaque and complex. As a result, maintaining the ever-important scientific value of transparency can be challenging and necessitates further development of practices and strategies to ensure its preservation.

Nevertheless, we agree that such concerns should not prevent scientific applications of LLMs. It is unrealistic to presume that LLMs won't be used because of the risks involved, and not permitting their use could do more harm than good: Given the current trend, if prohibited, they would likely be used covertly, exacerbating the already-worrying transparency issues. Certainly, we need to pursue a critical and not starry-eyed understanding of LLMs and maintain a clear-eyed assessment of the potential risks of use. However, there are ways of employing them that can improve the quality of science as long as the researcher is kept at the center of the process. LLMs are tools and, as such, must be carefully evaluated in their applications. This applies

to any tool, including the existing alternatives discussed by Bender et al., which, although optimized for specific scientific purposes, are not immune from mistakes and whose degree of reliability always needs careful scrutiny. At the end of the day, the responsibility falls upon the shoulders of the researchers who use the tools. It is, hence, crucial to establish principles and values that guide our decisions—whether one applies LLMs or any other method.

Ultimately, we mostly concur with Botvinick and Gershman: The impact of LLMs on the future practice of science cannot be fully predicted, but science is a humanistic and human enterprise and must remain so, motivating curbs to LLM use. Our perspective highlighted the normative aspects in terms of core values that should guide their use today, while Botvinick and Gershman seek to identify the principles and values for the future, deciding what should remain exclusively human even when AI becomes fully capable of performing every step of scientific inquiry as well as upholding values such as accountability, transparency, and fairness. The two perspectives complement each other in stimulating discussions about what should guide how we integrate AI into our scientific practices.

Response by Matthew M. Botvinick and Samuel J. Gershman

We see significant common ground across the other perspectives. We will focus here on one issue that gets to the heart of our perspective. Schulz et al. characterize LLMs as closer to collaborators than to tools. This raises critical issues of accountability, as pointed out by Bender et al. and Marelli et al. Some of these issues are currently being grappled with, while others will become more salient in the future as the technology advances. In particular, accountability is a fundamentally human concept: Humans are the only currently existing agents that are accountable in the sense that they have ultimate control over their own actions and voluntarily submit to a system that regulates these actions. Extending this concept to artificial agents would entail a profound shift in our attitudes, essentially requiring us to acknowledge the personhood of such agents.

This shift, if it ever happens, will have ramifications far beyond science. Policymakers are already starting to wrestle with the question of how accountability should operate in a world where AI systems are increasingly autonomous, and the issues can get quite complex. The difficulties can be bounded, however, in domains where humans are able to draw clear boundaries around what role they will permit AI systems to play. In science, we believe these boundaries should be firm and restrictive, limiting key decisions—and thus accountability—to human scientists.

Ultimately, we are interested in the limit case where the limits imposed on AI are sociological, moral, and juristic, rather than technological. To regard LLMs as genuine collaborators rather than sophisticated tools, we would need to acknowledge attributes of personhood that go far beyond the mere practice of science. Our view is that AI, no matter how intelligent, should remain a tool, because ceding personhood to artificial agents would have undesirable consequences. It's one thing for an AI scientist to tell us that there is a better way to fold proteins or design nuclear reactors, but it's quite another thing for it to tell us that it would rather be studying some other problem. It would also be quite a shock to be told by an AI scientist that it's solved an important problem but that it doesn't feel like trying to explain it to a human. As we argued in our perspective, the choices of what to study and which explanations count are irreducibly human.

Conclusion

We have presented four different perspectives centering around the question "how should the advancement of LLMs affect the practice of science?" Schulz et al. argued that "working with LLMs will not be fundamentally different from working with other collaborators, such as research assistants or doctoral students." Bender et al. described a suite of problems with using LLMs in scientific activity and argued that many uses of LLMs are "contrary to the norms of science." Marelli et al. called for "clear principles guiding the way this technology should be used," including transparency, accountability, and fairness. Finally, Botvinick and Gershman advocated that "two core aspects of scientific work should be reserved to human scientists," namely deciding on what problems to work on and that human understanding remains the goal of science.

A major point of contention in the context of doing science was the question of what systems like LLMs are capable of. Measuring and pinpointing these capabilities is a very hard problem. Typically, machine learning research does so through the use of benchmarks. It remains unclear, however, whether such benchmarks can be established for scientific workflows and whether success in such benchmarks would translate to meaningful scientific output in the real world. Yet, even though there was substantial disagreement, there were also important common themes. In particular, all parties emphasized the social nature of science and the importance of protecting scientific integrity and standards. In modern times, these core values are more important than ever before, and we—as a community—will have to continuously reevaluate how to protect them.

While most of the presented perspectives have focused on LLMs in their state, many of the raised points, e.g., transparency, accountability, and fairness, apply to the use of AI tools more broadly. We believe that based on the rapid development of this field, now is the time to establish norms about how we want to use such tools in contexts like optimal experimental design (85, 86), theory discovery (87, 88), and prediction (59, 89).

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. This work has been partially funded by the European Research Council (853489–DEXIM; 101087053–BraveNewWord), by the German Research Foundation (Deutsche Forschungsgemeinschaft) (2064/1–Project number 390727645), the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) (Tübingen AI Center, FKZ: 01IS18039A), and as part of the Excellence Strategy of the German Federal and State Governments.

1. Y. Bengio, R. Ducharme, P. Vincent, A neural probabilistic language model. *Adv. Neural Inf. Process. Syst.* **13**, 1137–1155 (2000).
2. D. Jurafsky, J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* (Pearson Prentice Hall, 2009).
3. T. Brown *et al*., Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
4. I. Drori *et al*., A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2123433119 (2022).
5. T. Kocmi, C. Federmann, "Large language models are state-of-the-art evaluators of translation quality" in *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, M. Nurminen *et al*., Eds. (European Association for Machine Translation, Tampere, Finland) (2023), pp. 193–203.
6. D. M. Katz, M. J. Bommarito, S. Gao, P. Arredondo, GPT-4 passes the bar exam. *Phil. Trans. R. Soc. A.* **382**, 20230254 (2024).
7. T. Eloundou, S. Manning, P. Mishkin, D. Rock, GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv [Preprint] (2023). https://arxiv.org/abs/2303.10130 (Accessed 19 February 2024).
8. E. Kasneci *et al*., Chatgpt for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **103**, 102274 (2023).
9. L. Messeri, M. Crockett, Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
10. R. Peres, M. Schreier, D. Schweidel, A. Sorescu, On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *Int. J. Res. Mark.* **40**, 269–275 (2023).
11. B. D. Lund, T. Wang, Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Libr. Hi Tech News* **40**, 26–29 (2023).
12. E. L. Hill-Yardin, M. R. Hutchinson, R. Laycock, S. J. Spencer, A Chat(GPT) about the future of scientific publishing. *Brain Behav. Immun.* **110**, 152–154 (2023).
13. H. Zheng, H. Zhan, ChatGPT in scientific writing: A cautionary tale. *Am. J. Med.* **136**, 725–726.e6 (2023).
14. B. D. Lund *et al*., ChatGPT and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *J. Assoc. Inf. Sci. Technol.* **74**, 570–581 (2023).
15. A. Birhane, A. Kasirzadeh, D. Leslie, S. Wachter, Science in the age of large language models. *Nat. Rev. Phys.* **5**, 277–280 (2023).
16. B. Fecher, M. Hebing, M. Laufer, J. Pohle, F. Sofsky, Friend or foe? Exploring the implications of large language models on the science system. arXiv [Preprint] (2023). https://arxiv.org/abs/2306.09928 (Accessed 19 February 2024).
17. C. Stokel-Walker, R. Van Noorden, What ChatGPT and generative AI mean for science. *Nature* **614**, 214–216 (2023).
18. R. Taylor *et al*., Galactica: A large language model for science. arXiv [Preprint] (2022). https://arxiv.org/abs/2211.09085 (Accessed 19 February 2024).
19. Embracing change and resetting expectations (2023). https://unlocked.microsoft.com/ai-anthology/terence-tao/. Accessed 4 September 2023.
20. W. D. Heaven, Why Meta's latest large language model survived only three days online. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/. Accessed 19 February 2024.
21. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2021), pp. 610–623.
22. K. Arkoudas, GPT-4 can't reason. arXiv [Preprint] (2023). https://arxiv.org/abs/2308.03762 (Accessed 4 July 2024).
23. F. Gilardi, M. Alizadeh, M. Kubli, ChatGPT outperforms crowd-workers for text-annotation tasks. arXiv [Preprint] (2023). https://arxiv.org/abs/2303.15056 (Accessed 19 February 2024).
24. D. U. Wulff, R. Mata, Automated jingle-jangle detection: Using embeddings to tackle taxonomic incommensurability. PsyArXiv [Preprint] (2023). https://doi.org/10.31234/osf.io/9h7aw (Accessed 19 February 2024).
25. D. Dillion, N. Tandon, Y. Gu, K. Gray, Can AI language models replace human participants? *Trends Cogn. Sci.* **27**, 597–600 (2023).
26. M. Hutson, Guinea pigbots. *Science* **381**, 121–123 (2023).
27. S. Trott, Large language models and the wisdom of small crowds. *Open Mind* **8**, 723–738 (2024).
28. B. Rozière *et al*., Code llama: Open foundation models for code. arXiv [Preprint] (2023). https://arxiv.org/abs/2308.12950 (Accessed 19 February 2024).
29. F. Sanmarchi *et al*., A step-by-step researcher's guide to the use of an AI-based transformer in epidemiology: An exploratory analysis of ChatGPT using the STROBE checklist for observational studies. *J. Public Heal.* **32** 1–36 (2023).
30. N. Dehouche, Plagiarism in the age of massive generative pre-trained transformers (GPT-3). *Ethics Sci. Environ. Polit.* **21**, 17–23 (2021).
31. P. P. Liang, C. Wu, L. P. Morency, R. Salakhutdinov, "Towards understanding and mitigating social biases in language models" in *International Conference on Machine Learning* (PMLR, 2021), pp. 6565–6576.
32. J. Coda-Forno *et al*., Inducing anxiety in large language models increases exploration and bias. arXiv [Preprint] (2023). https://arxiv.org/abs/2304.11111 (Accessed 19 February 2024).
33. B. Hutchinson *et al*., "Social biases in NLP models as barriers for persons with disabilities" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Eds. (Association for Computational Linguistics, 2020), pp. 5491–5501.
34. N. Carlini *et al*., "Extracting training data from large language models" in *30th USENIX Security Symposium (USENIX Security 21)* (USENIX Association, 2021), pp. 2633–2650.
35. The times sues openAI and microsoft over A.I. use of copyrighted work (2023). www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html. Accessed 22 January 2024.
36. M. R. King, A place for large language models in scientific publishing, apart from credited authorship. *Cell. Mol. Bioeng.* **16**, 95–98 (2023).
37. S. Herbold, A. Hautli-Janisz, U. Heuer, Z. Kikteva, A. Trautsch, AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays. arXiv [Preprint] (2023). https://arxiv.org/abs/2304.14276 (Accessed 19 February 2024).
38. R. A. Poldrack, T. Lu, G. Beguš, AI-assisted coding: Experiments with GPT-4. arXiv [Preprint] (2023). https://arxiv.org/abs/2304.13187 (Accessed 19 February 2024).
39. T. Goyal, J. J. Li, G. Durrett, News summarization and evaluation in the era of GPT-3. arXiv [Preprint] (2022). https://arxiv.org/abs/2209.12356 (Accessed 19 February 2024).
40. N. Yax, H. Anlló, S Palminteri, Studying and improving reasoning in humans and machines. *Commun. Psychol.* **2**, 51 (2024).
41. H. Touvron *et al*., Llama 2: Open foundation and fine-tuned chat models. arXiv [Preprint] (2023). https://arxiv.org/abs/2307.09288 (Accessed 19 February 2024).
42. A. Q. Jiang *et al*., Mixtral of experts. arXiv [Preprint] (2024). https://arxiv.org/abs/2401.04088 (Accessed 4 July 2024).
43. Towards a transparent AI future: The call for less regulatory hurdles on open-source AI in Europe (2003). https://laion.ai/blog/transparent-ai/. Accessed 22 October 2023.
44. W. Liang *et al*., Can large language models provide useful feedback on research papers? A large-scale empirical analysis. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.01783 (Accessed 19 February 2024).
45. E. M. Bender, A. Koller, "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*, D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Eds. (Association for Computational Linguistics, Online, 2020), pp. 5185–5198.
46. H. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (The MIT Press, 1992).
47. D. Wang, X. Wang, S. Lv, An overview of end-to-end automatic speech recognition. *Symmetry* **11**, 1018 (2019).
48. R. J. Hartsuiker, A. Moors, "On the automaticity of language processing" in *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*, H. J. Schmid, Ed. (American Psychological Association; De Gruyter Mouton, 2017).
49. R. M. Kinney *et al*., The semantic scholar open data platform. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2301.10140 (Accessed 19 February 2024).
50. S. Narayan, S. B. Cohen, M. Lapata, Ranking sentences for extractive summarization with reinforcement learning. arXiv [Preprint] (2018). https://arxiv.org/abs/1802.08636 (Accessed 19 February 2024).
51. D. Hodel, J. West, Response: Emergent analogical reasoning in large language models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2308.16118 (Accessed 19 February 2024).
52. P. Törnberg, D. Valeeva, J. Uitermark, C. Bail, Simulating social media using large language models to evaluate alternative news feed algorithms. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2310.05984 (Accessed 19 February 2024).
53. L. P. Argyle *et al*., Out of one, many: Using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).
54. F. Gilardi, M. Alizadeh, M. Kubli, ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2305016120 (2023).
55. G. Conroy, How ChatGPT and other AI tools could disrupt scientific publishing. *Nature* **622**, 234–236 (2023).
56. B. Latour, S. Woolgar, *Laboratory Life: The Construction of Scientific Facts* (Princeton University Press, 2013).
57. D. Partha, P. A. David, Toward a new economics of science. *Res. Policy* **23**, 487–521 (1994).
58. T. Amano *et al*., The manifold costs of being a non-native english speaker in science. *PLoS Biol.* **21**, e3002184 (2023).
59. J. Jumper *et al*., Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
60. F. Günther, L. Rinaldi, M. Marelli, Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Persp. Psychol. Sci.* **14**, 1006–1033 (2019).
61. S. Golchin, M. Surdeanu, Time travel in llms: Tracing data contamination in large language models. arXiv [Preprint] (2023). https://arxiv.org/abs/2308.08493 (Accessed 19 February 2024).
62. R. Li *et al*., Starcoder: May the source be with you! arXiv [Preprint] (2023). https://arxiv.org/abs/2305.06161 (Accessed 19 February 2024).
63. W. H. Walters, E. I. Wilder, Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci. Rep.* **13**, 14045 (2023).
64. J. Kocoń *et al*., ChatGPT: Jack of all trades, master of none. *Inf. Fusion* **99**, 101861 (2023).
65. H. Liu *et al*., Evaluating the logical reasoning ability of ChatGPT and GPT-4. arXiv [Preprint] (2023). https://arxiv.org/abs/2304.03439 (Accessed 19 February 2024).
66. E. Durmus *et al*., Towards measuring the representation of subjective global opinions in language models. arXiv [Preprint] (2023). https://arxiv.org/abs/2306.16388 (Accessed 19 February 2024).
67. M. Atari, M. J. Xue, P. S. Park, D. Blasi, J. Henrich, Which humans? PsyArXiv [Preprint] (2023). https://doi.org/10.31234/osf.io/5b26t (Accessed 19 February 2024).
68. S. Santurkar *et al*., Whose opinions do language models reflect? arXiv [Preprint] (2023). https://arxiv.org/abs/2303.17548 (Accessed 19 February 2024).
69. C. Flaherty, The peer review crisis (2023). https://www.insidehighered.com/news/2022/06/13/peer-review-crisis-creates-problems-journals-and-scholars. Accessed 30 October 2023.
70. M. Park, E. Leahey, R. J. Funk, Papers and patents are becoming less disruptive over time. *Nature* **613**, 138–144 (2023).
71. A. Davies *et al*., Advancing mathematics by guiding human intuition with AI. *Nature* **600**, 70–74 (2021).
72. J. Drescher, Out of DSM: Depathologizing homosexuality. *Behav. Sci.* **5**, 565–575 (2015).
73. N. H. Franco, Animal experiments in biomedical research: A historical perspective. *Animals* **3**, 238–273 (2013).

74. C. Debernardi, M. Seeber, M. Cattaneo, Thirty years of climate change research: A fine-grained analysis of geographical specialization. *Environ. Sci. Policy* **152**, 103663 (2024).
75. I. Gabriel, V. Ghazavi, "The challenge of value alignment: From fairer algorithms to AI safety" in *The Oxford Handbook of Digital Ethics*, C. Véliz, Ed. (Oxford University Press, 2022).
76. D. Silver *et al*., Mastering the game of go without human knowledge. *Nature* **550**, 354–359 (2017).
77. P. Lemos, N. Jeffrey, M. Cranmer, S. Ho, P. Battaglia, Rediscovering orbital mechanics with machine learning. *Mach. Learn. Sci. Technol.* **4**, 045002 (2023).
78. M. Botvinick, Have we lost our minds? (2023). https://medium.com/@matthew.botvinick/have-we-lost-our-minds-86d9125bd803. Accessed 30 October 2023.
79. N. Stiennon *et al*., Learning to summarize with human feedback. *Adv. Neural Inf. Process. Syst.* **33**, 3008–3021 (2020).
80. S. Longpre *et al*., "The flan collection: Designing data and methods for effective instruction tuning" in *Proceedings of the 40th International Conference on Machine Learning, ICML'23* (JMLR.org, 2023).
81. F. W. Taylor, *The Principles of Scientific Management* (Harper, 1913).
82. I. Grossmann *et al*., Ai and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
83. H. Zhao *et al*., Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.* **15**, 1–38 (2023).
84. H. Luo, L. Specia, From understanding to utilization: A survey on explainability for large language models. arXiv [Preprint] (2024). https://arxiv.org/abs/2401.12874 (Accessed 19 February 2024).
85. H. Wang *et al*., Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
86. Y. Strittmatter, B. Ji, S. Musslick, Automating the generation and documentation of experimental designs with natural language processing. PsyArXiv [Preprint] (2023). https://doi.org/10.31234/osf.io/2xk7t (Accessed 19 February 2024).
87. J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, T. L. Griffiths, Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **372**, 1209–1214 (2021).
88. S. Musslick, Y. Strittmatter, M. Dubova, Closed-loop scientific discovery in the behavioral sciences. PsyArXiv [Preprint] (2024). https://doi.org/10.31234/osf.io/c2ytb (Accessed 19 February 2024).
89. X. Luo *et al*., Large language models surpass human experts in predicting neuroscience results. arXiv [Preprint] (2024). https://arxiv.org/abs/2403.03230 (Accessed 19 February 2024).