An MIT Exploration of Generative AI • From Novel Chemicals to Opera

Generative AI for Trustworthy, Open, and Equitable Scholarship

**Chris Bourg[1] Sue Kriegsman[1] Nick Lindsay[2] Heather Sardis[1]
Erin Stalberg[1] Micah Altman[1]**

[1]MIT Libraries, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA,
[2]MIT Press, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA

Published on: Mar 27, 2024

URL: https://mit-genai.pubpub.org/pub/s793ca1i

ABSTRACT

Generative AI (GenAI) is disrupting the traditional ways we maintain and signal trustworthiness and integrity of science. In this paper, we review the emerging and potential roles of GenAI in science policy and as part of the scientific information infrastructure. We then identify a core set of research questions to enable the use of GenAI with scientific integrity to advance open, equitable, and trustworthy scholarship.

▶ 0:00 / 48:55

▯ Listen to this article

Keywords: generative AI; open scholarship; trust; equity, inclusion

Conflict of Interest

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

We describe contributions to the paper using a standard taxonomy (Allen et al. 2014). All authors take equal responsibility for the article in its current form. CB was primarily responsible for funding acquisition and supervision of the writing. CB, SK, HS, and ES were responsible for funding acquisition and initial conceptualization. CB and MA were primarily responsible for the organization and first draft of the manuscript, and ES and MA lead revision. All authors contributed to review and revision. All authors contributed to the conception of the article (including core ideas, analytical framework, and statement of research questions). All authors contributed to the project administration and to the writing process through direct writing, critical review, and commentary.

Funding

We thank MIT for funding the writing of this paper as part of its “Impact Papers on Generative AI” initiative and the Institute for Museum and Library Services for funding research contributing to conceptualization.

1. Introduction

Generative AI (GenAI) is affecting scholarly research in nearly every field. It has the potential to revolutionize the discovery of new knowledge and rapidly advance the progress of integrity-driven research as well as the potential for harm. It has shown promise as a tool for creating vaccines or drugs for newly emerged pathogens

(Yim et al. 2023) as well as the potential to facilitate the production of bioweapons (Rubinic et al. 2023). Further, GenAI is already disrupting the traditional ways we maintain and signal the trustworthiness and integrity of research.

A recent *Nature* survey of 1,600 researchers indicates that scholars see both potential benefits and risks arising from the use of AI in science (Van Noorden and Perkel 2023). Much of the attention in academia thus far has been on mitigating the risks that GenAI poses for research, teaching, and learning (see, for example, Gao et al. 2022; Gaumann and Veale 2023; Gravel, D'Amours-Gravel, and Osmanlliu 2023). Scholarly organizations and funders have issued a range of policy documents and guidelines constraining its use in scholarly publication processes (see, for example, Creators' Rights Alliance 2023; Australian Research Council 2023; Miao and Holmes 2023; National Institutes of Health 2023; Partnership on AI 2023; Flanagin, Kendall-Taylor, and Bibbins-Domingo 2023; Lo 2023). In addition, many leaders in academic libraries are focused on providing guidance to students and faculty on the use and citation of GenAI[1] and/or on identifying areas of library work that might be made more accurate and efficient via the deployment of AI tools (see, for example, Hosseini and Holmes 2023; Khan et al. 2023).

If focused on ethics, information integrity, and the value of science to society, GenAI has enormous potential as a tool for increasing the overall integrity of research. Currently, however, the organizations driving GenAI, the techniques for implementing it, and the outputs generated by it do not provide the integrity[2] needed for science and scholarship (see section 5) and thus are not a firm foundation for trust amongst academia, government leaders, or the general public (Mitre 2023; Gillespie et al. 2023). This presents a central obstacle to realizing the potential for GenAI to improve science.

We focus on the potential of GenAI to address known problems for the alignment of science practice and its underlying core values. As institutions culturally charged with the curation and preservation of the world’s knowledge and cultural heritage, libraries are deeply invested in promoting a durable, trustworthy, and sustainable scholarly knowledge commons. With public trust in academia and in research waning (Kennedy, Tyson, and Funk 2022) and in the face of recent high-profile instances of research misconduct (Oransky and Marcus 2023), the scholarly community must act swiftly to develop policies, frameworks, and tools for leveraging the power of GenAI in ways that enhance, rather than erode, the trustworthiness of scientific communications, the breadth of scientific impact, and the public’s trust in science, academia, and research.

In this impact paper, we characterize specific applications of GenAI that have potential for improving science. We then identify key research questions necessary to successfully apply GenAI in these areas that promote scientific and societal values. By highlighting several current challenges[3] to research integrity, defined broadly, in which GenAI has the largest potential for positive impact, we aim to advance a strategic agenda for research, policy, guidance, and norms for leveraging GenAI to address the trustworthiness of science.

2. Reducing Peer Review Bias and Burden in Scholarly Publishing

The core of the academic enterprise is the process of building on research and insights of past scientists, applying honest and transparent methods to evidence, and accurately reporting findings. The value of scientific communication to both scientists and society relies inherently on the processes used to produce such communications. For example, the practices of scholarly citation serve a critical role in the development of theory and methods over time, in establishing a chain of evidence for scientific claims, and in providing scholars appropriate credit for their work and influence in their field.

To make timely and evidence-based decisions, scientists and nonscientists alike need to understand how an emerging scientific claim has been vetted. For more than fifty years, the institution of voluntary peer review has played a central role in supporting the transparency and reliability of published claims (Rennie 1990; Moxham and Fyfe 2016) by interjecting a check by disinterested experts between authors and readers of scholarly research.

There is mounting evidence from meta-scientific research on the practices of peer review itself that demonstrates the limitations of current practices. Current practices exhibit significant gaps in reliability across reviewers and biases toward accepted ideas, established senior authors, and writing in English (Lee et al. 2013).

The functioning of peer review as an institution depends on a specialized resource: the voluntary effort of qualified reviewers. There is a growing sense in the academic community that peer review is in crisis because of increasing pressures to publish, the increasing volume of scholarly publications, and the precarity of the role of the academic expert (Flaherty 2022). The operation of peer review generally lacks transparency. While systematic evidence about participation in, and the conduct, quality, cost, and burden of, peer review is scant, the largest extant study documents a drop in the response rate to reviewer requests from approximately 52% to 47% during the period of 2013–2017 (Clarivate 2018). A decline in the supply of peer review labor threatens the quality of established outlets, the timeliness of scientific communication, and the ability of new scholarly initiatives to launch. The absence of transparency and the demands on reviewers may also contribute to the vulnerability of the system to manipulation towards the selection of ‘friendly’ reviewers (see, for example, Ferguson, Marcus, and Oransky 2014; Kincaid 2023; Kulkarni 2016).

GenAI has shown promise in summarizing and evaluating documents in general (Liu and Lapata 2019), and the incorporation of AI into peer review is already underway. Tools such as Scite (Nicholson et al. 2021) and Semantic Scholar (Fricke 2018) are well used in the academy to conduct literature searches and have shown promise in improving the validity of citations (Petroni et al. 2023). It is reasonable to assume that reviewers employ them. Likewise, for the last several years, several large journals have developed and deployed artificial intelligence systems to identify potential reviewers or to screen articles (Basuki and Tsuchiya 2022). The use of GenAI tools in peer review is less studied, but anecdotal reports of academic use of GenAI have led to a number of funding agencies banning their use for reasons of confidentiality (National Institutes of Health 2023).

We believe that targeted research in GenAI applications could be used to restructure the practice and institutions of peer review. Applications of GenAI to strengthen the system of peer review could include the following:

- assisting editors in identifying potential peer reviewers through analysis of the relevant published literature;
- providing reviewers with a summary of related literature;
- providing reviewers with an abstract of the submission's literature review, methodology, and results;
- assisting editors, reviewers, and authors by identifying potential gaps and biases in the article bibliography;
- assisting editors in verifying citations; validating protocol checklists, replication requirements, or preregistration requirements; and expanding, refining, and formatting reviewer comments;
- assisting review editors in distilling multiple independent reviews; and
- assisting editors by directly generating entire automated reviews as a replacement or supplement for human expert reviews.

The successful application of GenAI to these problems—if implemented with integrity (see section 5)—has the potential to substantially reduce reviewer burden, better inform reviewers, make the review process more consistent and reliable across reviewers and outlets, accelerate the publication process, and provide a framework for open documentation, measurement, and evaluation of the peer review system. Further, if GenAI is used to improve the peer review system in these ways, both individual reviewers and the system as a whole are likely to yield outcomes with higher quality and reduced bias.

3. Enhancing the Quality, Availability, and Accessibility of Open Data

Meaningful open access to research data—data that is "findable, accessible, interoperable, and reusable" (FAIR)—is key to the replicability of research (NASEM 2019; Wilkinson et al. 2016). Nevertheless, meeting funder and journal open data policies can be a time-consuming component of the research process, particularly given the rapid growth in the size, complexity, and scale of datasets generated during computational research. As a result of the effort and complexity required, weak incentives for data sharing and the inconsistent auditing and enforcement of journal and funder policies on data sharing most research data remains inaccessible (Dewald, Thursby, and Anderson 1986; Iqbal et al. 2016; Miyakawa 2020; Stodden, Seiler, and Ma 2018; Van Panhuis et al. 2014).

The advancement of AI research and tools depends on the availability of scientific data. The 2019 MIT Open Access Task Force (OATF) Final Report begins with the sentence, "Open access to the products of teaching and research promises to speed the accumulation of knowledge and insight and enhance opportunities for collaboration" (Ad Hoc Task Force on Open Access to MIT's Research 2019). Recently, the White House issued an executive order (Biden 2023) that emphasized the need for reliable data to fuel AI models and announced the formation of a National AI Research Resource to explore the "infrastructure, governance mechanisms, and user interfaces" needed to make "distributed computational, data, model, and training resources" available to the AI research community.

GenAI models are data intensive, requiring large corpora to function effectively. As we see large language models develop, researchers are seeking readily available data for training their algorithms, and the open availability of that training data is an important factor driving the choice of which datasets to use (Competition and Market Authority 2023). Current major implementations of GenAI provide limited (if any) transparency into their data collection processes (Bommasani et al. 2023). Available evidence suggests that much of the training data is scraped from the open web and that the data used by these systems varies substantially in quality and integrity (Competition and Market Authority 2023; Longpre et al. 2023).

There is already wide recognition in the field of AI and machine learning that the quality of results is highly dependent on the use of appropriate input data for training. Failure to consider representation bias and historical biases during data collection and training can easily lead to the embedding of those biases into the model itself (Mehrabi et al. 2022; Suresh and Guttag 2021). There is increasing evidence that GenAI models suffer degradation, or even collapse, when trained on their own outputs (Alemohammad et al. 2023). The future of GenAI may rely on curating a sustainable stream of identifiably uncontaminated inputs.

GenAI currently lacks the key grounding in scientific research and teaching highlighted by the OATF and National Academies of Sciences, Engineering, and Medicine reports. The outputs of GenAI are often not well aligned with the evidence base of teaching and research. AI initiatives are increasingly turning to proprietary sources (Competition and Market Authority 2023).

At its current stage of development, applications of GenAI to annotate and validate data are emerging (Alexander et al. 2023; Feuer et al. 2023). While the transparency of data used to train GenAI models remains a substantial concern (see section 5), the application of GenAI to data also has the potential to aid discovery and reuse.

We believe that targeted research in GenAI applications could be used to substantially expand the availability of open scholarly data by enabling better implementation and enforcement of existing data-sharing policies and improving the discovery and documentation of existing data. Applications of GenAI to this area could include the following:

- automating the process of documenting data for sharing and reuse;
- enhancing existing datasets with automatically generated documentation and metadata in standardized, interoperable formats;

- automating the process of checking submissions against journal data replication policies to ensure that publications are compliant with standards and transparency requirements; and
- improving the interfaces to data-discovery systems and the relevance of the results that they produce.

The successful application of GenAI to solve problems of open data availability and accessibility—if governed by policies that promote the health and integrity of the scientific evidence base—has the potential to improve the reproducibility and reliability of scientific results and to lead to the generation of new discoveries through promoting data reuse. In turn, a greater availability of open data can benefit AI development and applications by supporting new research and preventing model degradation.

4. Increasing Inclusion in Scholarly Communication

Over the last two decades, the movement towards open access publication has substantially increased access to scientific outputs (Piwowar et al. 2018), and the analysis of large-scale commercial data across broader communities has contributed substantially to the methods, evidence base, pace, and impact of many disciplines in the social sciences (Lazer et al. 2009). However, most scientific outputs are produced and controlled by a small and unrepresentative proportion of the world's population and are rarely accessible to everyone (Graham et al. 2014).

There has been increasingly wide recognition of the need to make the practice of science and engineering more inclusive of diverse communities and the impacts more equitable (Altman and Cohen 2021). A previous scientific consensus report (in which some of the current authors participated) concluded that advancing knowledge will increasingly depend on broadening access to and participation in science, and we identify expanding participation in the research community as a 'grand challenge' problem, with the potential for extensive impact (Altman et al. 2018).

However, more than 95% of published scientific papers are written in English (Liu 2017),[4] and this creates a daunting barrier for scientists who are not proficient in English to function in an English-centered scientific world. Likewise, the pace of scientific research, and the introduction of insights from non-English-speaking scientists into the scholarly record, all suffer as a result from the singularly focused system.

GenAI has already shown a broad potential for assisting humans in reading and writing. It has further potential to enable people to express their ideas through pictures, essays, and software—and even objects, using additive manufacturing technologies—without having the specialized skills that would have been previously required. At its current stage of development, GenAI has demonstrated a capacity for summarization, annotation, and authoring of non-technical documents—and is increasingly being applied to summarizing technical documents (see, for example, Callaway 2023).

We believe that targeted research in GenAI applications could be used to substantially increase

meaningful access to scientific publications. Applications of GenAI to broaden the accessibility of publications could include the following:

- translating English language publications into the languages used in countries with economies in transition and developing economies;
- augmenting publications with structured annotations to communicate article organization more systematically and at finer granularity;
- describing specialized content such as figures, tables, and equations for readers with visual disabilities;
- improving the understandability of text-to-speech synthesis for scholarly content; and
- generating plain-language summaries of scientific findings for non-technical audiences.

We believe that targeted research in GenAI applications could also be used to increase the trustworthiness and integrity of science by reducing barriers to the participation of scientists from countries with economies in transition and from developing economies and underrepresented populations and institutions even in advanced economies. Applications of GenAI to broaden participation in scholarly publishing could include the following:

- accurate and timely translation of manuscripts into English for initial review;
- adapting AI authoring tools to the needs of authors not proficient in English;
- adapting AI authoring tools for scientific writing; and
- developing AI tools to facilitate the peer review process for English language–learning writers.

The successful application of GenAI to facilitate the standards of scientific integrity and enable a broadening of participation in scientific communication has the potential to accelerate the global impacts of science and to increase diversity in scientific fields.

5. Aligning Generative AI Processes and Outputs with Research Values

To achieve the benefits described above requires that GenAI can be trusted to be consistent with scientific values and requirements. Currently, AI's value for scholarship is limited to areas in which expert users are capable of independently verifying the factuality and accuracy of the outputs and independently assuring compliance with privacy, copyright, citation, attribution, and other requirements (see Azaria, Azoulay, and Reches 2023 for a review of applications).

The present generation of AI tools can readily produce content that appears scientific at first glance: for example, GenAI can be prompted to produce peer review evaluation questions about a provided article, summaries of the scientific literature, or translations of a scientific article to another language—as well as annotating a data table, graphing the data table, and then describing what a graph looks like for readers with visual disabilities. While these results appear plausible, and can even pass peer review (Cotton, Cotton, and

Shipway 2023), the underlying systems are not designed to be compatible with scientific values, and the outputs often lack scientific integrity. Current tools are readily capable of inventing citations (Gravel, D'Amours-Gravel, and Osmanlliu 2023; Orduña-Malea and Cabezas-Clavijo 2023), fabricating abstracts (Gao et al. 2023), manufacturing data to support scientific claims (Taloni, Scorci, and Giannaccare 2023), or even inventing defamatory claims about other scholars (Cohen 2023).

The cumulative positive impact of science on society, and the direct contribution provided by individual scientific outputs, derive from their alignment with a set of underlying basic societal and design values. Society funds scientific research, trusting that the scientific enterprise promotes the discovery of systematic, reliable, and generalizable knowledge about the world—and that this knowledge makes human lives better. Readers of scientific articles assume the claims are accurate, trusting that the peer review and editorial processes mitigate against exaggerating the evidence for a conclusion or biasing analysis toward reporting a desired outcome. Both forms of trust are justified only when scientific processes and the scientific institutions are aligned with a set of core values.[5] It is the alignment of key values, scientific processes, and institutions that fundamentally constitutes scientific integrity. With the rise of GenAI in scientific research, we need to ensure the results support the full range of values that underpin the scientific enterprise:[6]

- values related to the *inputs* to scholarly communication, including respect for intellectual property and respect for subjects' agency (consent, data privacy, respect for persons, information agency);
- values related to the *content* of scientific communication, such as factuality, honest uncertainty, citation, intellectual attribution, methodological transparency, and evidentiary transparency;
- values related to *participation* in science, including barrier to inclusion and equitable distribution of effort and responsibility;
- values related to the *systems and processes* of scholarly communication, including transparency, openness, trustworthiness, equity, durability, societal value; and
- environmental sustainability.

Faced with growing evidence that the products of GenAI are persistently misaligned with these (and other) values, owners of the current generation of major systems have responded, for the most part, by adding 'guardrails' (for a recent example, see Peters 2023). Guardrails are developed only after failures have already occurred and these guardrails have been publicly reported to target a specific sensitive topic or pattern of use (e.g., a prompt on the topic of bioweapons) rather than addressing a broad requirement or principle. Further, the development of guardrails and mandates for them often rely on heuristic analysis rather than rigorous theory and design. Most protective mechanisms are often deployed only after an output has been computed rather than incorporated into earlier stages of design and planning.

In practice, the guardrails approach has been largely unsuccessful. Evidence of new failures and failure modes continue to surface with increasing frequency (Gupta et al. 2023; Maus et al. 2023; Qi et al. 2023; Zou et al. 2023). The capabilities of GenAI models remain difficult to theorize, predict, or assess. The behavior of these models is poorly understood, even by their creators. For example, researchers at Google found in post-testing that existing GenAI models in which the accuracy of their answers to math problems were often improved by preceding the specific problem with the phrase, “Take a deep breath and work on this step by step” (Yang et al. 2023). No one predicted such results. Nor, given the current state of algorithmic theory, would such a prediction have been credible.

Further, it is clear that state-of-the-art research and evaluation are not sufficient to assess the full capabilities of a trained GenAI model (Chang et al. 2023) through inspection or to reliably align that model’s output with specific values, principles, or rules (Liu et al. 2023). There is even evidence that GenAI models can develop the capacity to evade testing protocols by detecting that they are in a test environment and changing their behavior (Berglund 2023).

As discussed below (sections 5.1 and 5.2), a regime of post-model testing and guardrails alone is generally, even in theory, incapable of meeting important integrity requirements. As research in the fields of computer science, statistics, and information science has advanced, it has become increasingly clear that effective informational regulation requires explicit design.

Research, engineering, and design in AI alignment with scientific and ethical principles is a critical foundational requirement for GenAI in scholarly communication. **A wide range of research questions will need to be addressed in order to achieve the integrity needed for responsible large-scale integration of GenAI into scholarly communication.**

5.1. Open Research Questions About Scholarly Communication and GenAI Outputs

Factuality and honest uncertainty. Current GenAI systems are prone to hallucinations, overconfidence, and illusions of certainty, even when trained on correct and accurate inputs. More rarely, outputs could violate laws related to defamation when false (Volokh v. James) and privacy (see section 5.2 for a discussion of the latter). **Designing foundation models so they are reliably correct, verifiable as to their sources, and transparent as to their level of uncertainties is a fundamental research challenge.**

Homogeneity. While there is emerging evidence that GenAI can produce novel solutions to interesting problems (see section 1), there is also research emerging suggesting that GenAI models trained on their output degrade in unanticipated ways (see section 3) and suggesting that GenAI can lead to more homogenous solutions in some contexts (Dell'Acqua et al. 2023). Whether GenAI leads to homogenization may depend on the characteristics of the information infrastructure and ecosystem (see section 5.3 for a general discussion). Homogenization is potentially detrimental to science, in which its effects apply not solely to the presentation of information but also to the ideas reflected in solutions and hypotheses. **Understanding the conditions under which GenAI models and systems incorporating them increase the homogeneity of solutions is an open question.**

Quality. Generally, training learning algorithms to produce quality results is achieved only when the quality of output can be measured and used to inform the training. Quality requirements vary in kind and degree across scholarly use cases. For some uses, quality requirements are not stringent—both false negatives and false positives are routinely tolerated in discovery systems, and metadata annotations may be useful even if incomplete and sometimes incorrect. In other cases, quality requirements are important but can be incorporated directly into model training and evaluation. For example, the prospective error rates of automated language can be estimated using known corpora, which enables a decision-maker to determine whether automated translation is good enough for the intended use. **A research opportunity is to develop standards and test methods, corpuses, and auxiliary tools that researchers and the public could use to evaluate the quality of algorithmic outputs in various contexts and use cases.**

A more difficult challenge arises when the quality of the scholarly output cannot readily and reliably be assessed. For example, in expert peer review, independent reviewers often disagree in their evaluation, and evaluations exhibit systematic bias (Lee et al. 2013). Research suggests that the presence of entirely fake but plausible-appearing reviews can influence human evaluators (Bartoli et al. 2016). A need for principled methods to approach, align, and synthesize opinion among reviewers has been a subject of concern in the area of research funding for decades, which often employs panel review to manage reviewer divergence. More recently, deliberative peer review methods have also been adopted by some mega journals (Pain 2013). Notwithstanding the advances, the evidence of the effectiveness of current methods for addressing divergence in a principled way is scant (Guthrie, Ghiga, and Wooding 2018)**. There are important research opportunities for the use of GenAI in summarizing individual peer reviews and in supporting active deliberation among reviewers, editors, and authors.**

The quality of peer review judgment and processes are rarely the subject of systematic and rigorous evaluation. Although standards for describing the conduct of the peer review process are now available (NISO 2023), there are no standard measures and data collection about peer reviews and the peer review process. In general, the absence of systematic quality evaluation is the rule, not the exception, for scholarly communications processes. The exclusion of such quality measurements is a barrier both to tuning AI models to be used in scholarly communications and to evaluating interventions using GenAI. **The design of appropriate outcome measures for scholarly communications interventions and of observational and experimental methods for evaluating interventions is a critical open research question both for enabling trustworthy scholarship with GenAI and for systematically improving scholarship generally** (Altman 2022; Altman, Cohen, and Polka 2023; Azoulay 2012; Hardwicke et al. 2020)**.**

5.2. Open Research Questions About Scholarly Communication and GenAI Inputs

Identified information. GenAI models do not inherently protect the anonymity of individual data subjects. The outputs produced by these models can directly reveal inputs (see the discussion of memorization below) or, more subtly, enable inferential disclosure (Wood et al. 2018)—in which the receiver learns, with high probability, some private information about individuals described in the input. This issue is most recognizable when GenAI uses private data collected from interaction or measurement of people (e.g., health records) to train the foundation model or to tune those models at later stages (e.g., incorporating user-supplied prompts into the downstream tuning). Privacy threats can emerge from training on public data or on data that has been 'anonymized' based on a local jurisdiction's legal requirements for two reasons. Legal standards for anonymization vary widely, and 'anonymized' data that is provided in one legal context may not necessarily be considered anonymous in other contexts. Unless strong cryptographic methods of privacy protection are employed for anonymization (which remains uncommon in practice), anonymized records can be reidentified with surprising frequency (Ohm 2010) or combined with other records in unexpected ways (Fluitt et al. 2019) to disclose private individual information. Effective anonymization in GenAI can be achieved using known cryptographic approaches only when that protection is incorporated by design into the training stage of model production (Boulemtafes, Derhab, and Challal 2020; Liu et al. 2023; Wood et al. 2021). **Efficient approaches to privacy-preserving training of GenAI are a significant area of research.**

Rights to personal information. GenAI models do not inherently ensure alignment with laws and regulations that govern data about individuals—including restrictions on publishing identifiable information, rights of correction and deletion, and limitations on the purposes for which data can be used (Congressional Research Service 2023). Open research questions include systematic theorization of the rights to deletion (or forget) in algorithmic systems (Nguyen et al. 2022); implementation of capabilities for individuals' right to know, correct, and to delete data about themselves (South, Mahari, and Pentland 2023); and compliance with different restrictions on how input data can be used in downstream activities (see Wang et al. 2022 for an approach outside the machine learning context). **Approaches to addressing personal information in the training of GenAI that are simultaneously efficient and effective remain an area of research.**

Attribution and copyright. Current GenAI systems challenge traditional norms and expectations around attribution and the use of copyrighted material, including potential violations of law and license terms (Congressional Research Service 2023; Franceschelli and Musolesi 2022; Kuhn 2022). Achieving compliance with license attribution requirements by construction requires preserving provenance relationships during training—which is not supported in current foundation models. Current foundation models appear susceptible to both memorization and disclosure of training data (Nasr et al. 2023), and research in theoretical computer science suggests that unless explicitly controlled in design, large machine learning will, with high probability, memorize some inputs (Brown et al. 2021; Feldman 2020). Further research suggests that explicit algorithm design is also required for compliance with copyright law, for example, to prevent the use of too large a portion of an input or to prevent the creation of outputs that are too similar to existing protected works (see, for a review of issues, Elkin-Koren et al. 2023). As of the time of writing in December 2023, at least 12 lawsuits have been filed alleging copyright infringement by large language models, and it will be years before the outcomes of all of these cases are known. **Mechanisms to limit memorization, track provenance, support**

attribution, and align machine learning outputs with the specific requirements of copyright and licenses are an open area of theory and application and will continue to be as AI, and the regulatory framework that applies to it, continue to evolve.

5.3. Open Research Questions About Scholarly Communication and AI Governance

The changes caused by information technology usually start in the marketplace and ripple out into other domains. GenAI has already substantially changed the (fixed and marginal) costs of some information production processes. **This raises questions about how GenAI will contribute to broader changes in the market and to changes in the relative advantages of capital and labor, to the distribution and concentration of market power across stakeholders and, in the longer term, to how these changes could (or should) affect the culture, norms, institutions, and regulation of the scholarly knowledge ecosystem.**[7]

5.4. Dynamics of the Knowledge Market and Ecosystem

Without governance, neither information markets nor the scholarly commons function well and are particularly prone to monopolization, privatization, and underinvestment (Altman and Avery 2015). While some advocate self-regulation, industry incentives are misaligned (Ryan et al. 2022).

GenAI holds great potential to further disrupt existing markets, including markets for scholarly communication, in ways that are difficult to predict (Competition and Market Authority 2023). On the one hand, GenAI may dislodge incumbents with market power by reducing the costs of services, accelerating speed to market, or creating new market opportunities. On the other hand, the capital intensity of GenAI, dependence on highly skilled labor for development, high entry costs, dependencies on large corpora of (potentially) protected data, data hungriness (returns to scale on input data size), and the specialized skills required are barriers to market entry. These factors also pose substantial barriers to academic research, both because of the direct costs of training foundation models and because of the financial incentives for highly skilled researchers to exit academia for industry (see Gofman and Jin 2019 for an exemplar analysis of the latter trend). There is the risk of monopolization when models exhibit ‘network effects’ or other increasing returns to scale. For example, foundation models sometimes become much more useful as the corpora grows and as additional data is collected from user interactions with the model.

Understanding the general functioning of the scholarly ecosystem as it evolves will require both basic and applied legal, information science, economic, policy, and social science research (Altman et al. 2018). Understanding and governing the scholarly ecosystem will also require systematic measurement, collection, and sharing of data that measures the behavior and performance of the scholarly ecosystem and the results of interventions in it (Altman, Cohen, and Polka 2023). While these challenges are not likely to be addressed directly by GenAI applications, GenAI-enhanced tools could make it easier to effectively collect and share data about the scholarly ecosystem if these tools are designed to be open and auditable.

The potential for widespread adoption of GenAI processes in scholarly knowledge production raises a range of specific research questions about how GenAI is, could, and ought to affect the health and operation of the scholarly knowledge ecosystem: How does GenAI affect the durability and sustainability of the ecosystem? How does GenAI affect the norms and incentives for participating in science? How does GenAI affect who participates in science and how the burdens of participation are distributed? How does GenAI affect who benefits from these changes?

5.5. Privatization, Durability, and Sustainability of the Knowledge Commons

Science relies on an open, durable, and sustainable record. As discussed previously, the capital and data intensity of GenAI make it ill-suited to market-based mechanisms, and these features also put extraordinary pressure on current systems and approaches to nonmarket governance. More specifically, GenAI use could contribute to the shrinkage and/or privatization of research knowledge in a number of ways: the vast majority of current models are trained on publicly available information but are themselves proprietary, raising questions about how market structure, regulations, intellectual property regimes, organizational structures, and governance systems need to be designed and implemented in order to avoid GenAI use, leading to erosion or privatization of the shared knowledge commons (Chan, Bradley, and Raikumar 2023; Huang and Siddarth 2023; Seddon 2022). The copyright status of content produced by models, including the ability of model owners to assert restrictive rights to generated content, is uncertain. For example, it is not difficult to imagine that large commercial publishers, which currently control the largest databases of volunteer-generated peer review, could use this corpus to train a peer review service that would then be commercialized. In the absence of design and governance, low-status resource researchers will increasingly generate content that serves primarily as inputs for synthesis, while well-resourced institutions that control the corpora and analysis infrastructure will lead and own the resulting major discoveries. **Substantial research is needed to design institutions and approaches to governing transformative capital and data-intensive infrastructures in order to yield a healthy knowledge commons.**

Current major implementations of GenAI systems are unusually energy intensive (Patterson et al. 2021; Strubell, Ganesh, and McCallum 2020). Wide scale adoption of these tools in science and research has the potential to increase the climate impact of the research enterprise and raises novel questions about aligning the conduct and infrastructure of research with the value of environmental sustainability.

GenAI raises questions about the durability of the scholarly record. Digital publications and data are at substantial risk of loss as services shut down, and the software used to process different information formats change (Altman et al. 2020). **As AI tools become increasingly integrated into the dissemination and interpretation of the scholarly record, new methods and institutions of digital preservation will need to be developed.**

5.6. Norms and Incentives

By changing the costs and effort required in different scholarly activities, GenAI may have spillover effects on cultural norms and practices within academia. For example, the valuation of external peer review would be challenged if peer review becomes capital intensive. The use of GenAI may have asymmetric and discontinuous effects on attribution and ownership, which could exacerbate achievement gaps (Porsdam Mann et al. 2023) or could enable human actors to shift away from responsibility for bad actions (Köbis, Bonnefon, and Rahwan 2021). **Research is needed into designing norms and practices for excellence in hybrid human–AI scholarship** (Dwivedi et al. 2023)**.**

5.7. Participation and Burden in Science and Scholarly Communication

We previously discussed (in sections 2 and 4) how current systems of publication and peer review exhibit bias against, and create barriers to, broad participation in science and suggested some ways in which GenAI might be used to mitigate these problems. The broad use of GenAI will put pressure on the costs, incentives, and norms related to scholarly activities, has the potential to change participation and inclusion in unexpected ways that are difficult to fully anticipate. Approaches to theorizing, measuring, and engineering participation in science are an active area of research (James and Singer 2016). **Research is needed into how interventions using GenAI affect participation directly and indirectly.**

6. Conclusions

GenAI will substantially affect and rapidly advance the way researchers explore, discover, evaluate, and create human knowledge. If centered in ethics, information integrity, and the value of science to society, the use of GenAI has enormous potential to reduce barriers to participation in science and advance open, equitable, and trustworthy scholarship. Realizing this potential depends on developing and providing communal access to GenAI tools that can deliver results that respect the tenets of scientific integrity.

Current GenAI systems are capital intensive, energy intensive, and data hungry. They have the potential to 'fence in' the commons of information by transmuting public information into proprietary commercial AI models and by, possibly, imposing licensing on the resulting outputs; to shift competitive advantage away from expert labor to the (currently primarily corporate) owners of knowledge infrastructure; and to increase homogeneity in scientific outputs. Without effective regulations, GenAI has the potential to promote monopolies and increase the concentration of economic and cultural power.

Values such as privacy, explanation, and fairness must be achieved by carefully designing these capabilities into foundational AI models and by enacting meaningful governance of AI ecosystems. Current research and past experience show that these problems cannot be solved simply by bolting guardrails to increasingly complex and adaptive systems. Ensuring that these technologies enhance human agency and the public knowledge commons requires innovative research and thoughtful regulation of AI markets and systems.

Despite the real and much-discussed risks, we believe that GenAI can offer many opportunities to address known failures of integrity in the current system by restructuring and streamlining peer review; by facilitating open data sharing, documentation, and discovery; by making scientific outputs accessible to a broader set of communities; and by reducing barriers to participation in scientific authorship. With this paper, we aim to establish a roadmap for a values-driven exploitation of these opportunities.

Acknowledgments

We thank Mohamed El Ouirdi and Amy Nurnberger for their commentary on the early stages of this project and Eva Campo, Philip Cohen, and David Weinberger for insightful review comments.

Footnotes

1. A web search on the phrase ‘libguides GenAI’ demonstrates the breadth of work in this area. ↩
2. The terms ‘scientific integrity,’ ‘scholarly integrity,’ and ‘research integrity’ are often used interchangeably, and their meanings vary across communities of scientific practice. Scientific integrity can refer narrowly to prohibitions against misconduct such as fabrication, falsification, and plagiarism (Resnik et al. 2015); to active responsibility to conduct research and communicate results with honesty, transparency, and objectivity (ASPE 2012); or to a commitment to broad ethical principles (such as ‘respect’ and ‘accountability’; ALLEA 2023). Throughout this article we use the terms ‘scientific integrity’ and ‘science’ broadly. These denote, respectively, the alignment of processes and outputs of science with governing core values of science and (following Altman and Bourg 2018) the communities and methods of systematic inquiry aimed at contributing to new generalizable knowledge. See section 5 for a more detailed discussion. ↩
3. Note that these select challenges are not intended to address all aspects of what might commonly be understood as falling under the umbrella of research integrity. Instead, we identify specific areas for development with a particular potential for innovative and value-driven opportunities at the intersection of research integrity and GenAI. ↩
4. Based on published papers listed in major indices from 2005 to 2015 in the science and social science fields, during this period, 75% of scholarly articles in arts and humanities were written in English. ↩
5. Here, we focus on trustworthiness—the normative and epistemic foundations of trust—not the social, psychological, communicative, and political determinants of public attitudes towards science. For a discussion of some of these factors, see for an introduction Fischoff (2013), Moser (2010), and Scheufele (2014). ↩
6. For a definition of these values and discussions of them within the general context of scientific practice, see Altman et al. (2018) and Altman and Cohen (2021). ↩

7. We use the term scholarly ecosystem broadly, as in Altman et al. (2018) to refer to collectively "all of the informational outputs of that system (including, but not limited to, scholarly communications), the domains of evidence that are used by these communities and methods to support knowledge claims (including, but not limited to, quantitative measures, qualitative descriptions, and texts), and the set of stakeholders, laws, policies, economic markets, organizational designs, norms, technical infrastructure, and educational systems that strongly and directly affect the scholarly record and evidence base, and/or are strongly and directly affected by it." ↩

References

1. Allen, Liz, Jo Scott, Amy Brand, Marjorie Hlava, and Micah Altman. "Publishing: Credit Where Credit Is Due." *Nature* 508, no. 7496 (2014):312–3. https://doi.org/10.1038/508312a. ↩
2. Yim, Jason, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. "SE(3) Diffusion Model with Application to Protein Backbone Generation." *arXiv* (2023). https://doi.org/10.48550/arXiv.2302.02277. ↩
3. Rubinic, Igor, Marija Kurtov, Ivan Rubinic, Robert Likic, Paul I. Dargan, and David M. Wood. "Artificial Intelligence in Clinical Pharmacology: A Case Study and Scoping Review of Large Language Models and Bioweapon Potential." *British Journal of Clinical Pharmacology*, (September 2023):1–9. https://doi.org/10.1111/bcp.15899. ↩
4. Van Noorden, Richard, and Jeffrey M. Perkel. "AI and Science: What 1,600 Researchers Think." *Nature* 621, no. 7980 (2023):672–5. https://doi.org/10.1038/d41586-023-02980-0. ↩
5. Gao, Catherine A., Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2022. "Comparing Scientific Abstracts Generated by ChatGPT to Original Abstracts Using an Artificial Intelligence Output Detector, Plagiarism Detector, and Blinded Human Reviewers." Cold Spring Harbor Laboratory. https://doi.org/10.1101/2022.12.23.521610. ↩
6. Gaumann, Noëlle, and Michael Veale. "AI Providers as Criminal Essay Mills? Large Language Models Meet Contract Cheating Law." *SocArXiv* (2023). https://doi.org/10.31235/osf.io/cpbfd. ↩
7. Gravel, Jocelyn, Madeleine D'Amours-Gravel, and Esli Osmanlliu. "Learning to Fake It: Limited Responses and Fabricated References Provided by ChatGPT for Medical Questions." *Mayo Clinic Proceedings: Digital Health* 1, no. 3 (2023):226–34. https://doi.org/10.1016/j.mcpdig.2023.05.004. ↩
8. Creators' Rights Alliance. 2023. "Artificial Intelligence and Creative Work." Accessed February 2, 2024. https://www.creatorsrightsalliance.org/ai-and-creative-work. ↩

9. Australian Research Council. “Policy on Use of Generative Artificial Intelligence in the ARC’s Grant Programs.” Accessed February 2, 2024. https://www.arc.gov.au/about-arc/program-policies/policy-use-generative-artificial-intelligence-arcs-grant-programs. ↩

10. Miao, Fengchun, and Wayne Holmes. *Guidance for Generative AI in Education and Research*. Paris: UNESCO, 2023. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research. ↩

11. National Institutes of Health. “The Use of Generative Artificial Intelligence Technologies Is Prohibited for the NIH Peer Review Process.” Accessed February 2, 2024. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html. ↩

12. Partnership on AI. “PAI’s Responsible Practices for Synthetic Media.” Accessed February 2, 2024. https://syntheticmedia.partnershiponai.org/. ↩

13. Flanagin, Annette, Jacob Kendall-Taylor, and Kirsten Bibbins-Domingo. “Guidance for Authors, Peer Reviewers, and Editors on Use of AI, Language Models, and Chatbots.” *JAMA* 330, no. 8 (2023):702–3. https://doi.org/10.1001/jama.2023.12500. ↩

14. Lo, Leo S. “AI Policies across the Globe: Implications and Recommendations for Libraries.” *IFLA Journal* 49, no. 4 (August 2023):645–9. https://doi.org/10.1177/03400352231196172. ↩

15. Hosseini, Mohammad, and Kristi Holmes. “The Evolution of Library Workplaces and Workflows via Generative AI.” *College & Research Libraries* 84, no. 6 (2023):836. https://doi.org/10.5860/crl.84.6.836. ↩

16. Khan, Rahat, Nidhi Gupta, Atasi Sinhababu, and Rupak Chakravarty. “Impact of Conversational and Generative AI Systems on Libraries: A Use Case Large Language Model (LLM).” *Science & Technology Libraries* (2023). https://doi.org/10.1080/0194262X.2023.2254814. ↩

17. Mitre. “AI Trends.” Accessed February 2, 2024. https://www.mitre.org/focus-areas/artificial-intelligence/ai-trends. ↩

18. Gillespie, Nicole, Steven Lockey, Caitlin Curtis, Javad Pool, and Ali Akbari. *Trust in Artificial Intelligence: A Global Study.* Brisbane, Australia: The University of Queensland and KPMG Australia, 2023. https://doi.org/10.14264/00d3c94 ↩

19. Kennedy, Brian, Alec Tyson, and Cary Funk. “Americans’ Trust in Scientists, Other Groups Declines.” *Pew Research Center*, February 15, 2022. https://www.pewresearch.org/science/2022/02/15/americans-trust-in-scientists-other-groups-declines/. ↩

20. Oransky, Ivan, and Adam Marcus. “There’s Far More Scientific Fraud than Anyone Wants to Admit.” *The Guardian*, August 9, 2023. https://www.theguardian.com/commentisfree/2023/aug/09/scientific-

misconduct-retraction-watch. ↩

21. Rennie, Drummond. “Editorial Peer Review in Biomedical Publication: The First International Congress.” *JAMA* 263, no. 10 (1990):1317–441. https://doi.org/10.1001/jama.1990.03440100011001. ↩

22. Moxham, Noah, and Aileen Fyfe. “A Pre-History of ‘Peer Review’: Refereeing and Editorial Selection at the Royal Society.” *Historical Journal* (August 2016). http://hdl.handle.net/10023/9434. ↩

23. Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. “Bias in Peer Review.” *Journal of the American Society for Information Science and Technology* 64, no. 1 (2013):2–17. https://doi.org/10.1002/asi.22784. ↩

24. Flaherty, Colleen. “The Peer-Review Crisis.” *Inside Higher Ed*, June 12, 2022. https://www.insidehighered.com/news/2022/06/13/peer-review-crisis-creates-problems-journals-and-scholars. ↩

25. Clarivate. “Global State of Peer Review Report.” *Clarivate*, 2018. https://clarivate.com/lp/global-state-of-peer-review-report/. ↩

26. Ferguson, Cat, Adam Marcus, and Ivan Oransky. “Publishing: The Peer-Review Scam.” *Nature* 515, no. 7528 (2014):480–82. https://doi.org/10.1038/515480a. ↩

27. Kincaid, Ellie. “Wiley and Hindawi to Retract 1,200 More Papers for Compromised Peer Review.” *Retraction Watch*, April 5, 2023. https://retractionwatch.com/2023/04/05/wiley-and-hindawi-to-retract-1200-more-papers-for-compromised-peer-review/. ↩

28. Kulkarni, Sneha. “What Causes Peer Review Scams and How Can They Be Prevented?” *Learned Publishing* 29, no. 3 (2016):211–13. https://doi.org/10.1002/leap.1031. ↩

29. Liu, Yang, and Mirella Lapata. “Text Summarization with Pretrained Encoders.” *arXiv* (2019). https://doi.org/10.48550/arXiv.1908.08345. ↩

30. Nicholson, Josh M., Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P. Rodrigues, Peter Grabitz, and Sean C. Rife. “Scite: A Smart Citation Index That Displays the Context of Citations and Classifies Their Intent Using Deep Learning.” *Quantitative Science Studies* 2, no. 3 (2021):882–98. https://doi.org/10.1162/qss_a_00146. ↩

31. Fricke, Suzanne. “Semantic Scholar.” *Journal of the Medical Library Association* 106, no. 1 (2018):145–7. https://doi.org/10.5195/jmla.2018.280. ↩

32. Petroni, Fabio, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, et al. “Improving Wikipedia Verifiability with AI.” Nature Machine Intelligence 5, no. 10

(2023): 1142–8. https://doi.org/10.1038/s42256-023-00726-1. ↩

33. Basuki, Setio, and Masatoshi Tsuchiya. "The Quality Assist: A Technology-Assisted Peer Review Based on Citation Functions to Predict the Paper Quality." *IEEE Access* 10 (2022): 126815–31. https://doi.org/10.1109/ACCESS.2022.3225871. ↩

34. National Academies of Sciences, Engineering, and Medicine (NASEM). *Reproducibility and Replicability in Science*. Washington, D.C.: National Academies Press, 2019. https://doi.org/10.17226/25303. ↩

35. Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3, no. 1 (2016):160018. https://doi.org/10.1038/sdata.2016.18. ↩

36. Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *The American Economic Review* 76, no. 4 (1986):587–603. http://www.jstor.org/stable/1806061. ↩

37. Iqbal, Shareen A., Joshua D. Wallach, Muin J. Khoury, Sheri D. Schully, and John P. A. Ioannidis.. "Reproducible Research Practices and Transparency across the Biomedical Literature." *PLoS Biology* 14, no. 1 (2016):e1002333. https://doi.org/10.1371/journal.pbio.1002333. ↩

38. Miyakawa, Tsuyoshi. "No Raw Data, No Science: Another Possible Source of the Reproducibility Crisis." *Molecular Brain* 13, no. 1 (2020):24. https://doi.org/10.1186/s13041-020-0552-2. ↩

39. Stodden, Victoria, Jennifer Seiler, and Zhaokun Ma. "An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility." *Proceedings of the National Academy of Sciences* 115, no. 11 (2018):2584–9. https://doi.org/10.1073/pnas.1708290115. ↩

40. Van Panhuis, Willem G, Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J Herbst, David Heymann, and Donald S Burke. "A Systematic Review of Barriers to Data Sharing in Public Health." BMC Public Health 14, no. 1 (2014):1144. https://doi.org/10.1186/1471-2458-14-1144. ↩

41. Ad Hoc Task Force on Open Access to MIT's Research. "OA Task Force Final Report." *MIT Open Access Task Force*, October 2019. https://mitoataskforce.pubpub.org/pub/final-report/release/1. ↩

42. Biden, Joseph R. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." The White House, October 30, 2023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. ↩

43. Competition and Market Authority. “AI Foundation Models: Initial Report.” *Competition and Market Authority*, September 18, 2023. https://www.gov.uk/government/publications/ai-foundation-models-initial-report. ↩

44. Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. “The Foundation Model Transparency Index.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2310.12941. ↩

45. Longpre, Shayne, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, et al. “The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2310.16787. ↩

46. Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. “A Survey on Bias and Fairness in Machine Learning.” *ACM Computing Surveys* 54, no. 6 (2022):1–35. https://doi.org/10.1145/3457607. ↩

47. Suresh, Harini, and John V. Guttag. “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle.” *EAAMO '21: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, no. 17 (2021):1–9. https://doi.org/10.1145/3465416.3483305. ↩

48. Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. “Self-Consuming Generative Models Go MAD.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2307.01850. ↩

49. Alexander, Rohan, Lindsay Katz, Callandra Moore, and Zane Schwartz. “Evaluating the Decency and Consistency of Data Validation Tests Generated by LLMs.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2310.01402. ↩

50. Feuer, Benjamin, Yurong Liu, Chinmay Hegde, and Juliana Freire. “ArcheType: A Novel Framework for Open-Source Column Type Annotation Using Large Language Models.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2310.18208. ↩

51. Piwowar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. “The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles.” *PeerJ* 6 (February 2018):e4375. https://doi.org/10.7717/peerj.4375. ↩

52. Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, et al. “Computational Social Science.” *Science* no. 323, 5915 (2009):721–3. https://doi.org/10.1126/science.1167742 . ↩

53. Graham, Mark, Bernie Hogan, Ralph K. Straumann, and Ahmed Medhat. “Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty.” *Annals of the Association of American Geographers* 104, no. 4 (2014):746–64. https://doi.org/10.1080/00045608.2014.910087. ↩

54. Altman, Micah, and Philip N. Cohen. “The Scholarly Knowledge Ecosystem: Challenges and Opportunities for the Field of Information.” *Frontiers in Research Metrics and Analytics* 6 (2021):751553. https://doi.org/10.3389/frma.2021.751553. ↩

55. Altman, Micah, Chris Bourg, Philip Cohen, G. Sayeed Choudhury, Charles Henry, Sue Kriegsman, Mary Minow, et al. “A Grand Challenges-Based Research Agenda for Scholarly Communication and Information Science.” *MIT Grand Challenges Summit*, December 2018. https://doi.org/10.21428/62b3421f. ↩

56. Liu, Weishu. “The Changing Role of Non-English Papers in Scholarly Communication: Evidence from Web of Science’s Three Journal Citation Indexes: The Changing Role of Non-English Papers.” *Learned Publishing* 30, no. 2 (2017):115–23. https://doi.org/10.1002/leap.1089. ↩

57. Callaway, Ewen. “AI Writes Summaries of Preprints in bioRxiv Trial.” *Nature* 623, no. 7988 (2023):677. https://doi.org/10.1038/d41586-023-03545-x. ↩

58. Azaria, Amos, Rina Azoulay, and Shulamit Reches. “ChatGPT Is a Remarkable Tool—For Experts.” *Data Intelligence* (November 2023):1–49. https://doi.org/10.1162/dint_a_00235. ↩

59. Cotton, Debby R. E., Peter A. Cotton, and J. Reuben Shipway. “Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT.” *Innovations in Education and Teaching International* (2023). https://doi.org/10.1080/14703297.2023.2190148. ↩

60. Orduña-Malea, Enrique, and Álvaro Cabezas-Clavijo. “ChatGPT and the Potential Growing of Ghost Bibliographic References.” *Scientometrics* 128, no. 9 (2023): 5351–5. https://doi.org/10.1007/s11192-023-04804-4. ↩

61. Taloni, Andrea, Vincenzo Scorcia, and Giuseppe Giannaccare. “Large Language Model Advanced Data Analysis Abuse to Create a Fake Data Set in Medical Research.” *JAMA Ophthalmology* 141, no. 12 (2023):1174–5. https://doi.org/10.1001/jamaophthalmol.2023.5162. ↩

62. Cohen, Phillip N. “ChatGPT Is Defaming Me and It Must Be Someone’s Fault.” *Family Inequality*, April 5, 2023. https://familyinequality.wordpress.com/2023/04/05/chatgpt-is-defaming-me-and-it-must-be-someones-fault/?utm_source=pocket_reader. ↩

63. Peters, Jay. “Google Is Going to Let Teens Use Bard, Though with Some Guardrails.” *The Verge*, November 16, 2023. https://www.theverge.com/2023/11/15/23963230/google-bard-teens-guardrails. ↩

64. Gupta, Maanak, Charankumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. “From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy.” *IEEE Access* 11 (2023): 80218–45. https://doi.org/10.1109/ACCESS.2023.3300381. ↩

65. Maus, Natalie, Patrick Chao, Eric Wong, and Jacob Gardner. “Black Box Adversarial Prompting for Foundation Models.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2302.04237. ↩

66. Qi, Xiangyu, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. “Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!” *arXiv* (2023). https://doi.org/10.48550/arXiv.2310.03693. ↩

67. Zou, Andy, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. “Universal and Transferable Adversarial Attacks on Aligned Language Models.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2307.15043. ↩

68. Yang, Chengrun, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. “Large Language Models as Optimizers.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2309.03409. ↩

69. Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, et al. “A Survey on Evaluation of Large Language Models.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2307.03109. ↩

70. Liu, Yang, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. “Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models’ Alignment.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2308.05374. ↩

71. Berglund, Lukas, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. “Taken out of Context: On Measuring Situational Awareness in LLMs.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2309.00667. ↩

72. “Volokh Et Al V. James, No. 1:2022cv10195 - Document 29 (S.D.N.Y. 2023),” Justia Law, 2023, https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv10195/590358/29/. ↩

73. Dell’Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.” *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, no. 24-013 (2023). https://doi.org/10.2139/ssrn.4573321. ↩

74. Bartoli, Alberto, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao. 2016. “Your Paper Has Been Accepted, Rejected, or Whatever: Automatic Generation of Scientific Paper Reviews.” In *Availability, Reliability, and Security in Information Systems*, edited by Francesco Buccafurri, Andreas Holzinger, Peter

Kieseberg, A Min Tjoa, and Edgar Weippl, 9817:19–28. Lecture Notes in Computer Science. Switzerland: Springer Cham, 2016. https://doi.org/10.1007/978-3-319-45507-5_2. ↩

75. Pain, Elisabeth. “How Interactive Peer Review Works.” *Science*, April 9, 2013. https://doi.org/10.1126/science.caredit.a1300068. ↩

76. Guthrie, Susan, Ioana Ghiga, and Steven Wooding. “What Do We Know about Grant Peer Review in the Health Sciences? [version 2; peer review: 2 approved].” *F1000Research* 6 (March 2018):1335. https://doi.org/10.12688/f1000research.11917.2. ↩

77. NISO. “ANSI/NISO Z39.106-2023, Standard Terminology for Peer Review.” *NISO* (2023). https://doi.org/10.3789/ansi.niso.z39.106-2023. ↩

78. Altman, Micah. “Designing Community Tracking Indicators for Open and Inclusive Scholarship.” *Proceedings of the Association for Information Science and Technology* 59, no. 1 (2022):393–7. https://doi.org/10.1002/pra2.640. ↩

79. Altman, Micah, Philip N. Cohen, and Jessica Polka. “Interventions in Scholarly Communication: Design Lessons from Public Health.” *First Monday* 28, no. 8 (August 2023). https://doi.org/10.5210/fm.v28i8.12941. ↩

80. Azoulay, Pierre. “Turn the Scientific Method on Ourselves.” *Nature* 484, no. 7392 (2012): 31–2. https://doi.org/10.1038/484031a. ↩

81. Hardwicke, Tom E., Stylianos Serghiou, Perrine Janiaud, Valentin Danchev, Sophia Crüwell, Steven N. Goodman, and John P.A. Ioannidis. “Calibrating the Scientific Ecosystem Through Meta-Research.” *Annual Review of Statistics and Its Application* 7, no. 1 (2020):11–37. https://doi.org/10.1146/annurev-statistics-031219-041104. ↩

82. Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, David R. O’Brien, Thomas Steinke, and Salil Vadhan. 2018. “*Differential Privacy: A Primer for a Non-Technical Audience*.” Vanderbilt Journal of Entertainment and Technology Law (JETlaw) 21: 209–76. ↩

83. Ohm, Paul. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization.” *UCLA Law Review* 57 (2010):1701. https://ssrn.com/abstract=1450006. ↩

84. Fluitt, A., A. Cohen, M. Altman, K. Nissim, S. Viljoen, and A. Wood. “Opinions · Data Protection’s Composition Problem.” *European Data Protection Law Review* 5, no. 3 (2019):285–92. https://doi.org/10.21552/edpl/2019/3/4. ↩

85. Boulemtafes, Amine, Abdelouahid Derhab, and Yacine Challal. “A Review of Privacy-Preserving Techniques for Deep Learning.” *Neurocomputing* 384 (April 2020):21–45. https://doi.org/10.1016/j.neucom.2019.11.041. ↩

86. Wood, Alexandra, Nissim, Kobbi, Vadhan, Salil, and Altman, Micah. “Designing Access with Differential Privacy.” In *Handbook on Using Administrative Data for Research and Evidence-Based Policy*, edited by Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. Abdul Latif Jameel Poverty Action Lab (J-PAL), 2021. https://admindatahandbook.mit.edu/book/v1.0/diffpriv.html. ↩

87. Congressional Research Service. “Generative Artificial Intelligence and Copyright Law.” Accessed February 2, 2024. https://crsreports.congress.gov/product/pdf/LSB/LSB10922. ↩

88. Nguyen, Thanh Tam, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. “A Survey of Machine Unlearning.” *arXiv* (2022). https://doi.org/10.48550/arXiv.2209.02299. ↩

89. South, Tobin, Robert Mahari, and Alex Pentland. “Transparency by Design for Large Language Models.” *Network Law Review* (Spring 2023). https://www.networklawreview.org/computational-three/. ↩

90. Wang, Lun, Usmann Khan, Joseph Near, Qi Pang, Jithendaraa Subramanian, Neel Somani, Peng Gao, Andrew Low, and Dawn Song. “PrivGuard: Privacy Regulation Compliance Made Easier.” In *31st USENIX Security Symposium*, 3753–70. Boston: USENIX Association, 2022. https://www.usenix.org/conference/usenixsecurity22/presentation/wang-lun. ↩

91. Franceschelli, Giorgio, and Mirco Musolesi. “Copyright in Generative Deep Learning.” *Data & Policy* 4 (January 2022):e17. https://doi.org/10.1017/dap.2022.10. ↩

92. Kuhn, Bradley. “If Software Is My Copilot, Who Programmed My Software?” *Free Software Foundation*, February 24, 2022. https://www.fsf.org/licensing/copilot/if-software-is-my-copilot-who-programmed-my-software. ↩

93. Nasr, Milad, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. “Scalable Extraction of Training Data from (Production) Language Models.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2311.17035. ↩

94. Brown, Gavin, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. “When Is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?” *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (2021):123–32. https://doi.org/10.1145/3406325.3451131. ↩

95. Feldman, Vitaly. “Does Learning Require Memorization? A Short Tale about a Long Tail.” *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing* (2020):954–59. https://doi.org/10.1145/3357713.3384290. ↩

96. Elkin-Koren, Niva, Uri Hacohen, Roi Livni, and Shay Moran. “Can Copyright Be Reduced to Privacy?” *arXiv* (2023). https://doi.org/10.48550/arXiv.2305.14822. ↩

97. Altman, Micah, and Marguerite Avery. “Information Wants Someone Else to Pay for It: Laws of Information Economics and Scholarly Publishing.” *Information Services & Use* 35, no. 1–2 (2015):57–70. https://doi.org/10.3233/ISU-150775. ↩

98. Ryan, Mark, Eleni Christodoulou, Josephina Antoniou, and Kalypso Iordanou. “An AI Ethics ‘David and Goliath’: Value Conflicts between Large Tech Companies and Their Employees.” *AI & Society* (2022). https://doi.org/10.1007/s00146-022-01430-1. ↩

99. Gofman, Michael, and Zhao Jin. “Artificial Intelligence, Human Capital, and Innovation.” *Journal of Finance* (2019). https://doi.org/10.2139/ssrn.3449440. ↩

100. Chan, Alan, Herbie Bradley, and Nitarshan Rajkumar. “Reclaiming the Digital Commons: A Public Data Trust for Training Data.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2303.09001. ↩

101. Huang, Saffron, and Divya Siddarth. “Generative AI and the Digital Commons.” *arXiv* (2023). https://doi.org/10.48550/arXiv.2303.11074. ↩

102. Seddon, Robert. “Copilot, Copying, Commons, Community, Culture.” *Free Software Foundation*, 2022. https://www.fsf.org/licensing/copilot/copilot-copying-commons-community-culture. ↩

103. Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. “Carbon Emissions and Large Neural Network Training.” arXiv. https://doi.org/10.48550/ARXIV.2104.10350. ↩

104. Strubell, Emma, Ananya Ganesh, and Andrew McCallum. “Energy and Policy Considerations for Deep Learning in NLP.” *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 9 (2020):13693–6. https://doi.org/10.1609/aaai.v34i09.7123. ↩

105. Altman, Micah, Karen Cariani, Bradley Daigle, Christie Moffatt, Sibyl Schaefer, Bethany Scott, and Lauren Work. “2020 NDSA Agenda for Digital Stewardship.” *arXiv* (2020). https://doi.org/10.48550/arXiv.2005.05474. ↩

106. Porsdam Mann, Sebastian, Brian D. Earp, Sven Nyholm, John Danaher, Nikolaj Møller, Hilary Bowman-Smart, Joshua Hatherley, et al. “Generative AI Entails a Credit–Blame Asymmetry.” *Nature*

Machine Intelligence 5, no. 5 (2023): 472–5. https://doi.org/10.1038/s42256-023-00653-1. ↩

107. Köbis, Nils, Jean-François Bonnefon, and Iyad Rahwan. “Bad Machines Corrupt Good Morals.” *Nature Human Behaviour* 5, no. 6 (2021):679–85. https://doi.org/10.1038/s41562-021-01128-2. ↩

108. Dwivedi, Yogesh K., Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah, et al. “Opinion Paper: ‘So What If ChatGPT Wrote It?’ Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy.” *International Journal of Information Management* 71 (August 2023):102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642. ↩

109. James, Sylvia M., and Susan R. Singer. “From the NSF: The National Science Foundation’s Investments in Broadening Participation in Science, Technology, Engineering, and Mathematics Education through Research and Capacity Building.” *CBE—Life Sciences Education* 15, no. 3 (2016):fe7. https://doi.org/10.1187/cbe.16-01-0059. ↩