Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Responsible AI in Academic Research: A Systematic Review of Ethical Principles, Challenges, and Future Directions

Fredrick Michael Ogore

fogore@daystar.ac.ke

Daystar University Nairobi School of Science Engineering and Health https://orcid.org/0000-0001-7099-6761

Systematic Review

Keywords: Artificial intelligence, Research ethics, Research integrity, Algorithmic bias, Data governance, Generative AI, Systematic review

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11555829/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Artificial intelligence (AI) is now present in every stage of the academic research life cycle, yet the rules governing its use remain fragmented. Following PRISMA 2020, this systematic review used 34 structured search strings and citation chasing to search literature published between 2015 and 2025. Of 359 records identified, 124 studies and documents met the inclusion criteria: 40 empirical studies, 43 conceptual analyses or commentaries, 20 policy and guidance documents, 17 reviews and 4 methodological frameworks; 83% were published from 2023 onwards. Guidance converges on a narrow set of standards: AI cannot be an author, material use must be disclosed, and humans remain accountable. Disclosure thresholds, AI in peer review, AI-generated images, AI-text detectors and the regulatory treatment of research remain contested. Challenges fall into three clusters: algorithmic bias; data governance, privacy and consent; and transparency, reproducibility and integrity, where empirical studies document fabricated references, unreliable and biased detectors, and rapidly rising, mostly undisclosed AI use. Emerging responses include lifecycle-based guidance, structured documentation, tailored ethics review and interdisciplinary oversight. We propose an integrated framework linking five principles to five research stages and an agenda emphasizing context-sensitive ethics, public participation and international coordination.

1. Introduction

Artificial intelligence (AI) has moved from a specialized research object to a general-purpose research instrument. Machine learning models now predict protein structures with near-experimental accuracy (Jumper et al. 2021), accelerate hypothesis generation and experimental design across the natural sciences (Wang et al. 2023), and, since the public release of large language models (LLMs) such as ChatGPT in late 2022, assist researchers with literature searching, coding, drafting, translation, and editing. A survey of more than 1,600 researchers conducted by *Nature* found that a substantial share already used AI tools in their work and expected those tools to become central to research practice within a decade, while simultaneously voicing concern about misinformation, fraud, and bias (Van Noorden and Perkel 2023). Large-scale linguistic analyses indicate that LLM-modified text rose rapidly in scientific abstracts after 2022, most steeply in computer science (Liang et al. 2024a).

This diffusion has outpaced the evolution of common ethical standards. Traditional research-ethics frameworks, such as the principles of beneficence, justice, and respect for persons in the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979), were developed for investigator-participant relationships in which the researcher had control over the instruments of inquiry and could provide an explanation. AI violates these assumptions: models can be black boxes to their designers, trained on data collected without the knowledge of the people it describes, and capable of producing fluent but false content. At the same time, the multiplication of AI ethics guidelines, which had surpassed eighty by 2019 (Jobin et al. 2019), has resulted in a gap between abstract principles and operational practice, as noted by several authors (Mittelstadt 2019; Morley et al. 2020).

They have been mapped in existing reviews on general AI ethics principles (Jobin et al. 2019; Hagendorff 2020), on the use of generative AI in education (Kasneci et al. 2023) or on the particular question of AI authorship (Lund et al. 2023). Few studies collate the ethical consequences of using AI for academic research as a practice including responses of researchers, institutions, publishers, and regulators, areas of agreement and disagreement, and remaining unresolved challenges in one framework. Resnik and Hosseini (2025b) contend that current guidance on research ethics needs to be revised considerably to include AI as a research tool. But systematic evidence on current practices is dispersed across disciplines and document types.

This review addresses that gap. Drawing on both primary sources (empirical studies, audits, and surveys) and secondary sources (conceptual analyses, reviews, and policy and guidance documents), it asks four research questions (RQs):

- **RQ1.** What fundamental principles of ethical research are invoked for the use of AI in academic research, and how are they interpreted in this context?
- **RQ2.** What are the current practices of researchers, institutions, publishers, and regulators, and where do they converge or diverge?
- **RQ3.** What are the principal ethical challenges associated with AI-enabled research?
- **RQ4.** What strategies are emerging to address these challenges, and what future research is needed?

The review makes three contributions. First, it provides a detailed synthesis of ethical guidance and evidence from the social sciences, biomedicine, computer science, and research-integrity scholarship. Secondly it gives a structured comparison of the convergence and divergence of the most important sources of guidance. Third, it implies a prioritized research agenda and an integrated lifecycle framework. The remainder of the paper is organized as follows. Section 2 presents the conceptual underpinnings. The methodology for the review is presented in Section 3. Section 4 presents the results for RQ2–RQ4. Implications, future directions and constraints are discussed in Section 5 and Section 6 concludes.

2. Conceptual foundations: principles of ethical research in the age of AI

Historically, research ethics has been founded on principlism, which involves the application of a limited number of mid-level principles to organize moral reasoning (Beauchamp and Childress 2019). A comparable framework has been implemented in the field of AI ethics. Floridi and Cowls (2019) proposed a unified framework that would enhance the explicability of the four bioethical principles, while Jobin et al. (2019) identified transparency, justice and fairness, non-maleficence, responsibility, and privacy as the most frequently cited principles across 84 guidelines. The OECD Recommendation on AI (OECD 2019) and the UNESCO Recommendation on the Ethics of Artificial Intelligence, which was adopted by its 193 member states (UNESCO 2021), are intergovernmental instruments that codify

comparable commitments. Five principles were identified as the most pertinent to academic research (RQ1) for the purposes of this review.

2.1. Transparency and explicability

Research transparency has historically required methods to be described in a way that is sufficiently detailed to permit examination and replication. Three levels of transparency are obtained through the use of AI: explicability (the capacity to understand the logic behind model outputs), disclosure (the degree to which AI was used), and documentation (the data, model, parameters, and prompts). Technical tools for the third layer are provided by explainable AI research (Barredo Arrieta et al. 2020). For high-stakes decisions, Rudin (2019) argues that researchers should give priority to models that are naturally interpretable rather than post-hoc explanations of black-box systems. Additionally, reproducibility depends on openness: According to Haibe-Kains et al. (2020), high-profile AI studies may become effectively unverifiable due to the lack of shared code and model details.

2.2. Justice and fairness

The principle of justice relates to the fair allocation of the benefits and burdens of research. In AI-enabled research, this includes algorithmic fairness, which is the degree to which models trained on biased data systematically disadvantage groups based on race, gender, language, geography, or socioeconomic status (Mehrabi et al. 2021). The structural dimension of justice includes data extraction, training corpora knowledge system representation, and access to computational resources (Birhane 2021; Crawford 2021).

2.3. Respect for persons, autonomy, and privacy

Respect for individuals underpins informed consent and the protection of individual rights. These mechanisms are often contested in big-data and AI research, as data is often repurposed, scraped from public platforms or combined in ways that participants could not have foreseen (Metcalf and Crawford 2016). Rocher et al. (2019) estimated that 99.98% of Americans can be re-identified correctly in any dataset using 15 demographic attributes, suggesting that de-identification offers weaker protection than is commonly assumed.

2.4. Accountability and responsibility

Responsibility implies that identifiable individuals should be accountable for the conduct and outcomes of research. Publishers and ethics bodies have clarified that artificial intelligence systems are not able to meet authorship criteria because they lack the ability to assume moral or legal obligations (COPE 2023; ICMJE 2024). Accountability requires that decisions based on AI are contestable and responsibility is not spread over developers, vendors and users (Raji et al. 2020).

2.5. Research integrity and beneficence

When generative models produce plausible but fictitious content, the fabrication, falsification, and plagiarism, honesty, and rigor that constitute research integrity are directly engaged. AI-enabled research should also be beneficent and non-maleficent, that is, it should offer a net benefit and avoid foreseeable harm, including environmental costs associated with training large models (Strubell et al. 2019; Bender et al. 2021).

Table 1 summarizes how these principles are operationalized into specific duties across the research life-cycle. It is also the analytical lens used during data extraction (Section 3.5).

Table 1
Operationalization of core ethical principles for AI-enabled academic research.

Principle	Core question for AI-enabled research	Illustrative operational obligations
Transparency / explicability	Can others understand what AI did, and verify it?	Disclose AI tools, versions, and purposes; share code, data, and prompts where lawful; prefer interpretable models in high-stakes settings
Justice / fairness	Who benefits, who is burdened, and is the model biased?	Audit training data and outputs for subgroup performance; report demographic composition; ensure equitable access and benefit sharing
Respect for persons / privacy,	Were data subjects' rights and expectations respected?	Obtain appropriate consent or ethics waiver; minimise and secure data; assess re-identification risk; avoid uploading confidential data to third-party tools
Accountability	Who answers for AI-assisted outputs and decisions?	Only humans as authors; named responsibility, for verification; documented human oversight; mechanisms for contestation and correction
Integrity / beneficence	Are outputs genuine and is the research net beneficial?	Verify every AI-generated claim and citation; prohibit AI-fabricated data and images; weigh environmental and social costs

Source: Authors' synthesis based on Beauchamp and Childress (2019), Floridi and Cowls (2019), Jobin et al. (2019), and UNESCO (2021).

3. Methodology

3.1. Review design and protocol

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement (Page et al. 2021). Because the evidence base is heterogeneous, combining empirical, conceptual, and normative documents, a qualitative evidence synthesis rather than a meta-analysis was performed. The protocol, specifying research questions, search strings, eligibility criteria, and synthesis approach, was defined.

3.2. Information sources and search strategy

Searches were conducted on 14 October 2026, and eligibility was restricted to documents published between 1 January 2015 and 31 December 2025. Because direct exports from subscription databases (Scopus, Web of Science) were not available to the author, 34 structured search strings were executed through a general scholarly web search that indexes publisher platforms (including ScienceDirect, SpringerLink, Wiley Online Library, SAGE, and Taylor & Francis), PubMed/PMC, the ACM Digital Library, and the arXiv and medRxiv preprint servers. Each string combined a technology concept (e.g., artificial intelligence, machine learning, large language models, ChatGPT, generative AI) with an ethics or integrity concept (e.g., ethics, integrity, bias, privacy, transparency, disclosure, authorship) and a research-practice context (e.g., scientific research, scholarly publishing, peer review, ethics committees), and the set of strings was designed to cover every theme in the review protocol. Every record returned was logged (Supplementary Table S1, sheets Search_Log and Screening_Log). Additional records were identified through backward citation chasing from included studies and through known-item checks of guidance issued by COPE, ICMJE, WAME, major journals and publishers, the European Union, and UNESCO. The generic Boolean structure underlying the strings, suitable for re-execution in bibliographic databases, was:

("artificial intelligence" OR "machine learning" OR "deep learning" OR "large language model*" OR "generative AI" OR ChatGPT) AND (ethic* OR integrity OR "responsible research" OR bias OR fairness OR transparen* OR privacy OR accountab*) AND ("academic research" OR "scientific research" OR "scholarly publishing" OR "research practice*" OR "peer review" OR authorship OR "research ethics")

3.3. Eligibility criteria

Eligibility criteria were defined using an adapted Population–Concept–Context framework suited to scoping and qualitative reviews (Table 2).

Table 2
Eligibility criteria.

Dimension	Inclusion	Exclusion	Code
Publication date	1 January 2015–31 December 2025	Published before 2015 or after 2025	E1
Population / actors	Researchers, research institutions, ethics committees/IRBs, funders, publishers, editors, reviewers, regulators	Consumers, industry product teams, clinicians in routine care, students in classroom assessment	E2
Concept	Ethical, integrity, or governance aspects of using AI to conduct, report, review, or govern research	Purely technical or how-to work without ethical or integrity analysis	E3
Context	Academic and scientific research in any discipline, including scholarly publishing and peer review	General AI ethics without a research-practice component; clinical deployment; industry practice; teaching and learning	E2
Document type	Peer-reviewed empirical, review, and conceptual articles; peer-reviewed commentaries; preprints from recognised servers reporting original data or analysis; official guidance from publishers, editorial bodies, governments, and intergovernmental organisations	News, blogs, vendor pages, social media, library how-to guides, slides, AI-generated summaries	E4
Verifiability	Venue and authorship verifiable	Unrefereed uploads without venue; case notes; unindexed outlets	E5
Language	English	Other languages	E2
Insufficient information	–	Title, date, or venue could not be verified (reported as awaiting classification)	AWC

3.4. Selection process and quality appraisal

All records were recorded in a screening log (Supplementary Table S1) and duplicates removed, including alternative copies, versions and aggregator listings of the same work. Titles and abstracts were screened against the inclusion criteria and potentially eligible reports were retrieved in full text. Each exclusion was assigned a reason code (E1–E5; Table 2) and records for which bibliographic details could not be verified were listed as awaiting classification, as recommended by PRISMA 2020. One reviewer, aided by AI, conducted the screening and data extraction. Entries extracted from titles and abstracts rather than full texts are indicated in the workbook. The methodological quality of empirical studies was assessed with the Mixed Methods Appraisal Tool (MMAT) and policy and guidance documents with the AACODS checklist (authority, accuracy, coverage, objectivity, date, and significance).

Quality ratings were used to inform the weight of evidence in the synthesis, but were not used for studies exclusion. The selection process is described in the PRISMA flow diagram (Fig. 1).

3.5. Data extraction and synthesis

Data were extracted by the author using a standardized extraction form for bibliographic details, document type, discipline, stage of the research lifecycle addressed, ethical principles invoked (Table 1), challenges identified, key findings, and mitigation measures proposed. The extraction table of all included studies is completed and available in Supplementary Table S1 (sheet Included_Studies). Data were synthesized using thematic synthesis (Thomas and Harden 2008) involving line-by-line coding of findings, organization of codes into descriptive themes, and development of analytical themes mapped to RQ2–RQ4. For RQ2, guidance documents were further compared against a common set of provisions to identify convergence and divergence.

4. Results

4.1. Characteristics of the included evidence

The searches identified 314 records, and a further 45 were identified through other methods (Fig. 1). After removal of 63 duplicates, 296 records were screened by title and abstract, of which 139 were excluded: 37 fell outside the 2015–2025 window (predominantly items published in 2026), 55 were non-scholarly sources, and 47 addressed AI ethics outside academic research practice. Of the 157 reports assessed in full, 5 lacked an ethical or integrity focus, 9 were neither peer-reviewed nor authoritative guidance, and 19 could not be verified and are reported as awaiting classification. In total, 124 studies and documents were included in the synthesis.

The included evidence is recent and mostly normative (Fig. 2). Only 21 items (17%) were published between 2015 and 2022, whereas the remaining 103 (83%) appeared in 2023–2025, after the public release of ChatGPT. The most common are conceptual analyses and commentaries (n = 43) and empirical studies (n = 40), followed by policy and guidance documents (n = 20), reviews (n = 17), and methodological frameworks (n = 4). Empirical work has concentrated on four designs: corpus and linguistic analyses to estimate the prevalence of AI-generated text (Gray 2024; Liang et al. 2024a; Kobak et al. 2025); experimental tests of generative-AI outputs and detectors (Walters and Wilder 2023; Weber-Wulff et al. 2023; Hartung et al. 2024); analyses of journal and institutional policies (Ganjavi et al. 2024; Lund and Naheem 2024); and researcher surveys (Van Noorden and Perkel 2023). Publication policy (n = 28 studies), research integrity (n = 26), and authorship and disclosure (n = 19) were the most frequently addressed themes, whereas global equity (n = 4), environmental sustainability (n = 2), and public participation (n = 1) were the least studied. Medicine and the biomedical sciences, computer science, and research-ethics scholarship contributed most of the evidence.

4.2. Current practices: convergence and divergence (RQ2)

4.2.1. Areas of convergence

Guidance issued by editorial bodies and publishers following the emergence of ChatGPT converged rapidly on three norms. First, **AI tools cannot be authors**. After ChatGPT was listed as a co-author on several papers and preprints (Stokel-Walker 2023), Nature (2023) and *Science* (Thorp 2023) declared that LLMs do not satisfy authorship criteria, a position subsequently adopted by COPE (2023), WAME (Zielinski et al. 2023), ICMJE (2024), the JAMA Network (Flanagin et al. 2023), a group of bioethics journal editors (Kaebnick et al. 2023), and Elsevier (2023). Analyses of author instructions confirm the pattern: by 2023, 87% of the top 100 journals but only 24% of the top 100 publishers provided guidance on generative AI, and nearly all that did prohibited AI authorship (Ganjavi et al. 2024; see also Lund and Naheem 2024). This is because authorship entails responsibility and a non-agent cannot be responsible. Second, any substantive use of AI should be disclosed, generally either in the methods or acknowledgments section or as a dedicated statement. Third, the accuracy, originality, and integrity of all content, including any produced with AI assistance, are the sole responsibility of human researchers (Hosseini et al. 2023).

There is also convergence at the level of general AI governance. The commitment to transparency, fairness, oversight by humans, privacy and accountability, as reflected in the OECD (2019) principles, the UNESCO (2021) Recommendation, and the NIST AI Risk Management Framework (NIST 2023), echoes the principle-level consensus documented by Jobin et al. (2019).

4.2.2. Areas of divergence

There is consensus, under which practice varies in at least five respects (Table 3). (i) Disclosure thresholds vary: some policies exclude basic language editing and grammar correction, while while others require disclosure of any generative AI use. The normative literature is split. Resnik and Hosseini (2025a) argue for mandatory disclosure when AI plays a substantive role in shaping content, data, or analysis, but optional disclosure for routine language editing. Disclosure of writing assistance is at the discretion of the author, according to Hosseini et al. (2025). Science prohibited AI-generated text outright at first, before adopting a disclosed use policy. That shows how quickly the mood has changed. (ii) Peer review is viewed in an inconsistent way. Some organizations prohibit reviewers from uploading manuscripts to generative-AI tools due to confidentiality concerns (Elsevier 2023; ICMJE 2024), and researchers have suggested carefully limited uses to combat reviewer burnout (Hosseini and Horbach 2023). At the same time, this usage is already prevalent, with estimates that LLMs significantly altered 6.5–16.9% of peer-review text at major AI conferences in 2023–2024 (Liang et al. 2024b), AI-assisted reviews were associated with higher scores and acceptance rates (Latona et al. 2025), and existing policies have been criticized as inadequate to safeguard review integrity (Mollaki 2024). (iii) Publishers generally do not permit use of AI-generated images and data unless they are part of the research design, but institutional rules vary. (iv) Some institutions employ detection tools despite evidence of their unreliability (Weber-Wulff et al. 2023; Elkhatat et al. 2023) and bias against non-native English writers, misclassifying more than half of TOEFL essays as AI-generated (Liang et al. 2023). (v) Regulatory scope

varies The EU Artificial Intelligence Act (European Union 2024) significantly excludes artificial intelligence systems developed and used exclusively for scientific research and development from its scope, deferring research governance to institutional and professional self-regulation.

Table 3
Convergence and divergence in selected guidance on AI use in research and scholarly publishing.

Source	AI as author	Disclosure required	AI in peer review	Distinctive provisions
COPE (2023)	Not permitted	Yes	Not specified in position statement	Authors fully responsible for content, including AI-produced parts
ICMJE (2024)	Not permitted	Yes (cover letter and manuscript)	Reviewers must not upload manuscripts where confidentiality cannot be assured	Authors should be able to assert no plagiarism in AI-produced text and images
WAME (Zielinski et al. 2023)	Not permitted	Yes	Disclosure required if used	Editors need appropriate tools to detect AI-generated content
Nature (2023)	Not permitted	Yes (methods or acknowledgements)	Restricted	Generally does not permit AI-generated images
Science (Thorp 2023)	Not permitted	Initially banned text; later disclosed use permitted	Restricted	Policy revised within the same year
Elsevier (2023)	Not permitted	Yes (dedicated declaration)	Reviewers and editors must not upload manuscripts	AI-generated images not permitted unless part of research design
JAMA Network (Flanagin et al. 2023)	Not permitted	Yes (acknowledgements or methods)	Must not enter manuscripts into tools that breach confidentiality	Authors responsible for accuracy of AI-generated content
European Commission (2024)	n/a	Yes (substantial use)	Avoid substantial use in peer review and evaluation	Living guidelines for researchers, research organisations, and funders
EU AI Act (European Union 2024)	n/a	Transparency duties for certain systems	n/a	Exclusion for AI used solely for scientific R&D
UNESCO (2021); OECD (2019); NIST (2023)	n/a	Transparency as general principle	n/a	Principle-based and voluntary; not research-specific

Note: Policies evolve rapidly; entries reflect the cited versions and should be checked against current publisher and regulator guidance.

4.2.3. Researcher and institutional practice

From an empirical point of view, practice has led policy. Surveys indicate widespread acceptance of generative AI for writing and coding and literature work, with uncertainty about disclosure and concern about accuracy (Van Noorden and Perkel 2023; Elsevier 2024). Linguistic analyses suggest that use is massively under-disclosed: about 1% of papers published in 2023 contained marker vocabulary of LLM assistance (Gray 2024); LLM-modified content made up as much as 17.5% of computer-science abstracts by early 2024 (Liang et al. 2024a); and at least 13.5% of 2024 PubMed abstracts were processed with LLMs (Kobak et al. 2025), well beyond rates of formal disclosure. Often, institutional policies are delegated to the faculty or departmental level, resulting in inconsistency across and within universities. Research ethics committees and IRBs do not typically have the mandate or expertise to review AI-specific risks including dataset provenance, model bias, and downstream societal harms, especially for research that uses existing or publicly available data and is thus outside of traditional human-subjects review (Metcalf and Crawford 2016; Friesen et al. 2021; Ferretti et al. 2021; Bouhouita-Guermech et al. 2023); in response, the U.S. Secretary's Advisory Committee on Human Research Protections released specific considerations for IRB review of AI research (SACHRP 2022). The NeurIPS broader-impact requirement provided a natural experiment in self-governance in computer science, but analyses found statements of uneven depth that rarely engaged with concrete harms (Nanayakkara et al. 2021; Ashurst et al. 2022; Liu et al. 2022).

4.3. Principal ethical challenges (RQ3)

Thematic synthesis identified three dominant challenge clusters and two cross-cutting concerns. Table 4 summarises representative evidence and associated mitigation strategies.

4.3.1. Algorithmic bias and representational harm

Bias can enter AI-enabled research at every stage: through unrepresentative sampling, historically biased labels, proxy variables, model design, and deployment context (Mehrabi et al. 2021). Landmark audits illustrate the stakes. Buolamwini and Gebru (2018) found error rates of up to 34.7% for darker-skinned women in commercial gender classifiers, compared with 0.8% for lighter-skinned men. Obermeyer et al. (2019) showed that a widely used health-risk algorithm, which used healthcare costs as a proxy for need, substantially under-identified Black patients for additional care. When such tools are used as research instruments, bias becomes a validity problem as well as an ethical one, because findings may systematically misrepresent under-served populations. Generative models introduce further representational harms by reproducing stereotypes embedded in web-scale training data (Bender et al. 2021; Weidinger et al. 2022), and machine-learning research itself encodes values, such as performance and generalisation, that can crowd out considerations of justice (Birhane et al. 2022). Purely technical fairness fixes might fail if they abstract away from social context," warn Selbst et al. (2019). A more recent concern arises when LLMs are used as substitutes for human participants or as research assistants in the social and behavioral sciences: model outputs reflect the skewed populations and

viewpoints represented in training data, so findings may not generalize to underrepresented groups (Dillion et al. 2023; Grossmann et al. 2023; Bail 2024).

4.3.2. Data governance, privacy, and consent

The provenance, consent status and licensing of the large datasets underpinning AI research is often unclear. Web scraping data provokes questions as to whether data being publicly available is equivalent to consent to be used in research (Metcalf and Crawford 2016; Zook et al. 2017). The re-identification risks presented by Rocher et al. (2019) call into question the adequacy of anonymization as a safeguard. A new challenge unique to generative AI is the transmission of confidential content, such as unpublished manuscripts, participant data and grant proposals, to third-party services whose data-retention and training practices may be opaque. This is the basis for the confidentiality provisions in peer-review policies (ICMJE 2024). Robust data governance that follows the FAIR principles of findability, accessibility, interoperability and reusability (Wilkinson et al. 2016) must therefore be reconciled with privacy and data-protection obligations.

4.3.3. Transparency, reproducibility, and research integrity

Insufficient reporting and ambiguity threaten reproducibility. Gundersen and Kjensmo (2018) found that only a small portion of AI papers include enough information for replication. Kapoor and Narayanan (2023) discovered data leakage in 294 papers across 17 scientific fields using machine learning which inflated the reported performance. Generative AI brings new integrity risks. LLMs “hallucinate” plausible but false statements and references. Walters and Wilder (2023) found that 55% of citations generated by GPT-3.5 and 18% by GPT-4 were fabricated (see also Alkaissi and McFarlane 2023). Comparable results were reported in medicine, where nearly half of the references in ChatGPT-generated medical content were fabricated and most of the remainder inaccurate (Bhattacharyya et al. 2023), and hallucination rates for references proposed for systematic reviews reached 39.6% for GPT-3.5, 28.6% for GPT-4, and 91.4% for Bard (Chelli et al. 2024). Gao et al. (2023) showed that blinded human reviewers identified only 68% of ChatGPT-generated scientific abstracts, while misclassifying 14% of genuine abstracts as generated, which highlights the potential for undetectable paper-mill content (Kendall and Teixeira da Silva 2024). Image integrity is similarly threatened: generative models can produce realistic, fabricated research images (Gu et al. 2022), and expert pathologists could not reliably distinguish AI-generated from genuine histological images (Hartung et al. 2024). Messeri and Crockett (2024) further warn of “illusions of understanding," in which reliance on AI leads scientists to believe they understand more than they do and narrows the range of questions pursued, a concern echoed in expert assessments of how LLMs should and should not change scientific practice (Binz et al. 2025).

4.3.4. Cross-cutting concerns: accountability and sustainability

Two other concerns keep coming up. First, diffuse accountability: when AI is used to assist with data analysis or writing, it may not be obvious who is responsible for errors among researchers, tool designers, and institutions (Raji et al. 2020). Second, sustainability of the environment and of knowledge.

Training and running large models consume large amounts of energy and carbon (Strubell et al. 2019). Moreover, recursively training on AI-generated material can degrade model quality (Shumailov et al. 2024). These raise concerns about the long-term integrity of the scholarly record if synthetic text proliferates in the literature.

Table 4
Principal ethical challenges, representative evidence, and mitigation strategies.

Challenge	Mechanism	Representative evidence	Mitigation strategies
Algorithmic bias	Unrepresentative data, biased labels, proxy variables	Buolamwini and Gebru (2018); Obermeyer et al. (2019); Mehrabi et al. (2021)	Subgroup performance reporting; datasheets; bias audits; participatory design
Privacy and consent	Data repurposing, scraping, re-identification, third-party upload	Metcalf and Crawford (2016); Rocher et al. (2019)	Data minimisation; privacy-preserving methods; updated consent and REC review; secure local tools
Reproducibility	Incomplete reporting; leakage; proprietary models	Gundersen and Kjensmo (2018); Haibe-Kains et al. (2020); Kapoor and Narayanan (2023)	Code and data sharing; model cards; AI reporting guidelines; version pinning
Integrity of outputs	Hallucination, fabricated citations, synthetic text and images	Walters and Wilder (2023); Chelli et al. (2024); Gao et al. (2023); Hartung et al. (2024)	Mandatory human verification; disclosure statements; editorial checks
Detection and fairness	Unreliable, biased AI-text detectors	Weber-Wulff et al. (2023); Elkhatat et al. (2023); Liang et al. (2023)	Do not rely on detectors as sole evidence; due-process safeguards
Accountability	Diffused responsibility across actors	Raji et al. (2020); COPE (2023)	Human authorship; documented oversight; contribution statements (e.g., CRediT)
Sustainability and epistemics	Energy use; model collapse; narrowing of inquiry	Strubell et al. (2019); Shumailov et al. (2024); Messeri and Crockett (2024)	Report compute; proportionality assessment; methodological pluralism

4.4. Emerging strategies (RQ4)

4.4.1. Ethical standards and guidelines

The first response has been the rapid production of standards. Beyond the publication-ethics guidance in Table 3, research-specific frameworks are emerging that map ethical obligations to research stages rather than stating abstract principles (Resnik and Hosseini 2025b). Bockting et al. (2023) proposed

“living guidelines” overseen by an independent scientific body, an approach subsequently reflected in the European Commission's living guidelines on the responsible use of generative AI in research (European Commission 2024). Critics note that principles alone lack enforcement mechanisms and may function as ethics-washing (Mittelstadt 2019; Munn 2023). The literature therefore increasingly emphasises translating principles into tools, processes, and accountability structures (Morley et al. 2020; Whittlestone et al. 2019).

4.4.2. Promoting openness through documentation and reporting

Structured documentation is the most mature practical strategy. *Datasheets for datasets* record motivation, composition, collection processes, and recommended uses (Gebru et al. 2021), while *model cards* report intended use and disaggregated performance (Mitchell et al. 2019), and Data Cards extend this approach to dataset lifecycles (Pushkarna et al. 2022). Uptake remains partial, however: an analysis of 32,000 model cards found that sections on limitations, bias, and environmental impact were frequently missing or minimal (Liang et al. 2024c). In biomedicine, AI-specific reporting guidelines, namely CONSORT-AI and SPIRIT-AI for trials (Liu et al. 2020; Cruz Rivera et al. 2020) and TRIPOD + AI for prediction models (Collins et al. 2024), operationalise transparency for journals and reviewers. Broader open-science infrastructures, including code and data sharing mandates and the Transparency and Openness Promotion (TOP) guidelines (Nosek et al. 2015), provide complementary mechanisms.

4.4.3. Adapting ethics review and governance

Several authors propose extending the remit of RECs and IRBs to include AI-specific risk assessment, covering dataset provenance, bias, and downstream harm, even where research does not involve direct human participants (Friesen et al. 2021; Ferretti et al. 2021; Bouhouita-Guermech et al. 2023; SACHRP 2022). Risk-management frameworks such as the NIST AI RMF, organised around the functions *govern, map, measure*, and *manage* (NIST 2023), and accountability auditing frameworks (Raji et al. 2020) offer templates that institutions can adapt.

4.4.4. Interdisciplinary approaches and capacity building

A recurring recommendation is to embed ethical reflection within research teams through interdisciplinary collaboration between computer scientists, domain experts, ethicists, legal scholars, and social scientists, consistent with frameworks of responsible research and innovation that emphasise anticipation, reflexivity, inclusion, and responsiveness (Stilgoe et al. 2013). AI-literacy training for researchers, reviewers, and ethics-committee members is widely advocated so that they can recognise model limitations, verify outputs, and apply disclosure norms consistently (Dwivedi et al. 2023; van Dis et al. 2023; Binz et al. 2025).

5. Discussion

5.1. Synthesis: from principles to practice

At the heart of this review is a persistent principle-practice gap. While there are agreements on high-level values and a limited set of publication rules (e.g., no AI authorship, disclosure, and human accountability), these rules apply to the visible end of the research lifecycle (writing and publishing) rather than the upstream stages where many ethical risks exist (data collection and modelling). What distinguishes these differing disclosure thresholds, peer review, and detection rates is not merely institutional variation but real normative uncertainty about where assistance ends and authorship begins.

The five principles in Table 1 are considered cross-cutting obligations that are applicable at any stage of the research lifecycle. These principles are supported by the governance enablers identified in Section 4.4. The framework's value lies in shifting ethical attention upstream: bias audits and consent review at the data stage and documentation and verification at the modelling stage, rather than disclosure alone at the publication stage.

5.2. Future research directions

Building on the gaps identified, we propose a research agenda organised around three priorities (Table 5).

New ethical frameworks appropriate to AI's complexity. Principist frameworks are troubled by distributed agency, emergent model behavior and the scale of data reuse. Future work should develop and empirically test context-sensitive frameworks, such as relational approaches to algorithmic justice (Birhane 2021) and graded disclosure taxonomies that distinguish editing, analytic, and generative uses of AI. Notably, there are few empirical studies examining how researchers actually use AI and how disclosure affects reader trust and reviewer judgment.

Enhanced public participation. Research which uses population-scale data and which produces knowledge with societal implications deserves public input into its governance. AI research contexts under-study mechanisms such as citizens' juries, community advisory boards and participatory dataset design. An important future direction is to evaluate their impacts on research legitimacy, data quality and equity, especially for communities that have been historically excluded from or harmed by research.

International collaboration towards global guidelines. There are still major gaps left by the research exemption under the EU AI Act, by the voluntary nature of the UNESCO and OECD instruments, and by the Global North focus of the guidance. To counter regulatory arbitrage and ensure guidance reflects diverse research contexts, comparable policy research, harmonization of publisher and funder mandates, and the inclusion of Global South institutions in standard-setting are required.

Table 5
Proposed agenda for future research.

Direction	Priority research questions	Suggested methods
Context-sensitive ethical frameworks	How should disclosure be graded by type of AI use? When does AI assistance undermine authorship or originality? How can relational justice be operationalised in research design?	Normative analysis; Delphi studies; vignette experiments; framework validation studies
Empirical research on practice	How do researchers actually use and disclose AI? What is the prevalence of undisclosed use and AI-fabricated content?	Large-scale surveys; bibliometric and linguistic analysis; audit studies
Public participation	Which participatory mechanisms improve the legitimacy and equity of AI research? How can affected communities shape dataset governance?	Participatory action research; deliberative forums; comparative case studies
International coordination	How do national and publisher policies diverge, and with what consequences? How can Global South perspectives be embedded in standards?	Comparative policy analysis; multi-stakeholder consensus processes
Ethics review and tooling	How effective are adapted REC/IRB processes, datasheets, and reporting guidelines at reducing harm?	Pre–post evaluations; randomised trials of reporting interventions; implementation science

5.3. Implications for practice

For researchers, the findings suggest that it is their responsibility to fact-check every claim and reference made by AI; to document AI use throughout the lifecycle, not only in writing; and not to upload confidential or personal data to external tools without permission. For institutions, they call for clear, harmonized policies on AI use, AI-competent ethics review and training. For journals and publishers, they recommend moving toward graded disclosure templates, adopting AI-specific reporting guidelines, and not relying only on unreliable detectors. They cite the need to fill the governance gap left by research exemptions and support for international coordination for funders and regulators.

5.4. Limitations

There are several limitations to this review. First, the searches were conducted using a general scholarly web search and not directly exported from Scopus or Web of Science, and each search string only yielded the highest-ranked results; the corpus is therefore a structured and reproducible, but non-exhaustive sample, and relevant studies may have been missed. All search strings are reported to enable the search to be replicated in bibliographic databases and are included in the screening log. Second, screening and data extraction were performed by a single reviewer assisted by AI, and some entries of extraction are based on titles and abstracts, which are flagged in Supplementary Table S1 for verification. Thirdly, 19 records are yet to be classified as they could not be verified. Fourth, the field is moving fast, policies may have changed since the search date, and the restriction to English-language

sources limits generalizability to non-Anglophone and Global South contexts. Finally, the heterogeneity of included documents precluded quantitative synthesis.

6. Conclusion

AI is transforming the way academic research is conceived, conducted, reported and evaluated. This systematic review indicates that research community has come to swift agreement on a minimal set of norms: no AI authorship, disclosure of use, and human accountability, but remains divided on disclosure thresholds, peer review, detection, and regulatory scope. The central ethical issues are algorithmic bias, data governance and risks to transparency and integrity. Responses are evolving through standards, documentation, adapted ethics review and interdisciplinary collaboration. The integrated framework we propose shifts the locus of ethical attention upstream across the research lifecycle. Future work should focus on context-sensitive ethical frameworks, significant public participation, and international collaboration toward globally coherent guidance. Ethics are not an external constraint on AI-enabled research; they are a condition of its credibility.

Declarations

Author Contributions: Conceptualization, F.M.O.; methodology, F.M.O.; investigation, F.M.O.; formal analysis, F.M.O.; writing—original draft preparation, F.M.O.; writing—review and editing, F.M.O.; visualization, F.M.O. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The search strings, the complete screening log (359 records with decisions and reasons), and the data-extraction table for the 124 included studies are provided as Supplementary Table S1 (Excel workbook).

Acknowledgments: During the preparation of this work, the author used QuillBot and Paperpal to rephrase, polish, draft and edit the text and to assist with structuring. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Conflicts of Interest: The author declares no conflicts of interest.

Supplementary Materials: The following supporting information is available with this article: Table S1: Systematic review evidence workbook (Excel), containing the search strings, the complete screening log (359 records with decisions and reasons), and the data-extraction table for the 124 included studies.

References

1. Alkaissi, H., and S. I. McFarlane. 2023. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* 15: e35179. https://doi.org/10.7759/cureus.35179
2. Ashurst, C., E. Hine, P. Sedille, and A. Carlier. 2022. AI ethics statements: Analysis and lessons learnt from NeurIPS broader impact statements. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://arxiv.org/abs/2111.01705
3. Bail, C. A. 2024. Can generative AI improve social science? *Proceedings of the National Academy of Sciences* 121: e2314021121. https://doi.org/10.1073/pnas.2314021121
4. Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
5. Beauchamp, T. L., and J. F. Childress. 2019. *Principles of biomedical ethics* (8th ed.). Oxford University Press.
6. Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, pp. 610–623. https://doi.org/10.1145/3442188.3445922
7. Bhattacharyya, M., V. M. Miller, D. Bhattacharyya, and L. E. Miller. 2023. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus* 15: e39238. https://doi.org/10.7759/cureus.39238
8. Binz, M., S. Alaniz, A. Roskies, B. Aczel, C. T. Bergstrom, C. Allen, et al. 2025. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences* 122: e2401227121. https://doi.org/10.1073/pnas.2401227121
9. Birhane, A. 2021. Algorithmic injustice: A relational ethics approach. *Patterns* 2: 100205. https://doi.org/10.1016/j.patter.2021.100205
10. Birhane, A., P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, pp. 173–184. https://doi.org/10.1145/3531146.3533083
11. Bockting, C. L., E. A. M. van Dis, R. van Rooij, W. Zuidema, and J. Bollen. 2023. Living guidelines for generative AI — why scientists must oversee its use. *Nature* 622: 693–696. https://doi.org/10.1038/d41586-023-03266-1
12. Bouhouita-Guermech, S., P. Gogognon, and J.-C. Bélisle-Pipon. 2023. Specific challenges posed by artificial intelligence in research ethics. *Frontiers in Artificial Intelligence* 6: 1149082. https://doi.org/10.3389/frai.2023.1149082
13. Buolamwini, J., and T. Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research* 81: 77–91.

14. Chelli, M., J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, et al. 2024. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research* 26: e53164. https://doi.org/10.2196/53164
15. Collins, G. S., K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, et al. 2024. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385: e078378. https://doi.org/10.1136/bmj-2023-078378
16. Committee on Publication Ethics (COPE). 2023. *Authorship and AI tools: COPE position statement*. https://publicationethics.org/cope-position-statements/ai-author
17. Crawford, K. 2021. *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
18. Cruz Rivera, S., X. Liu, A.-W. Chan, A. K. Denniston, M. J. Calvert, and the SPIRIT-AI and CONSORT-AI Working Group. 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine* 26: 1351–1363. https://doi.org/10.1038/s41591-020-1037-7
19. Dillion, D., N. Tandon, Y. Gu, and K. Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* 27: 597–600. https://doi.org/10.1016/j.tics.2023.04.008
20. Dwivedi, Y. K., N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, et al. 2023. "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: 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
21. Elkhatat, A. M., K. Elsaid, and S. Almeer. 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity* 19: 17. https://doi.org/10.1007/s40979-023-00140-5
22. Elsevier. 2023. *The use of generative AI and AI-assisted technologies in writing for Elsevier* [Publishing ethics policy]. https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals
23. Elsevier. 2024. *Insights 2024: Attitudes toward AI* [Survey report]. Elsevier.
24. European Commission. 2024. *Living guidelines on the responsible use of generative AI in research*. Directorate-General for Research and Innovation, ERA Forum.
25. European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L series, 12 July 2024.
26. Ferretti, A., M. Ienca, M. Sheehan, A. Blasimme, E. S. Dove, B. Farsides, et al. 2021. Ethics review of big data research: What should stay and what should be reformed? *BMC Medical Ethics* 22: 51. https://doi.org/10.1186/s12910-021-00616-4
27. Flanagin, A., J. Kendall-Taylor, and K. Bibbins-Domingo. 2023. Guidance for authors, peer reviewers, and editors on use of AI, language models, and chatbots. *JAMA* 330: 702–703.

https://doi.org/10.1001/jama.2023.12500

28. Floridi, L., and J. Cowls. 2019. A unified framework of five principles for AI in society. *Harvard Data Science Review* 1. https://doi.org/10.1162/99608f92.8cd550d1
29. Friesen, P., R. Douglas-Jones, M. Marks, R. Pierce, K. Fletcher, A. Mishra, et al. 2021. Governing AI-driven health research: Are IRBs up to the task? *Ethics & Human Research* 43: 35–42. https://doi.org/10.1002/eahr.500085
30. Ganjavi, C., M. B. Eppler, A. Pekcan, B. Biedermann, A. Abreu, G. S. Collins, I. S. Gill, and G. E. Cacciamani. 2024. Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: Bibliometric analysis. *BMJ* 384: e077192. https://doi.org/10.1136/bmj-2023-077192
31. Gao, C. A., F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, and A. T. Pearson. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine* 6: 75. https://doi.org/10.1038/s41746-023-00819-6
32. Gebru, T., J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. 2021. Datasheets for datasets. *Communications of the ACM* 64: 86–92. https://doi.org/10.1145/3458723
33. Gray, A. 2024. *ChatGPT "contamination": Estimating the prevalence of LLMs in the scholarly literature* [Preprint]. arXiv:2403.16887. https://doi.org/10.48550/arXiv.2403.16887
34. Grossmann, I., M. Feinberg, D. C. Parker, N. A. Christakis, P. E. Tetlock, and W. A. Cunningham. 2023. AI and the transformation of social science research. *Science* 380: 1108–1109. https://doi.org/10.1126/science.adi1778
35. Gu, J., X. Wang, C. Li, J. Zhao, W. Fu, G. Liang, and J. Qiu. 2022. AI-enabled image fraud in scientific publications. *Patterns* 3: 100511. https://doi.org/10.1016/j.patter.2022.100511
36. Gundersen, O. E., and S. Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence* 32: 1644–1651. https://doi.org/10.1609/aaai.v32i1.11503
37. Hagendorff, T. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines* 30: 99–120. https://doi.org/10.1007/s11023-020-09517-8
38. Haibe-Kains, B., G. A. Adam, A. Hosny, F. Khodakarami, Massive Analysis Quality Control (MAQC) Society Board of Directors, L. Waldron, et al. 2020. Transparency and reproducibility in artificial intelligence. *Nature* 586: E14–E16. https://doi.org/10.1038/s41586-020-2766-y
39. Hartung, J., S. Reuter, V. A. Kulow, M. Fähling, C. Spreckelsen, and R. Mrowka. 2024. Experts fail to reliably detect AI-generated histological data. *Scientific Reports* 14. https://doi.org/10.1038/s41598-024-73913-8
40. Hosseini, M., and S. P. J. M. Horbach. 2023. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review* 8: 4. https://doi.org/10.1186/s41073-023-00133-5

41. Hosseini, M., D. B. Resnik, and K. Holmes. 2023. The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Research Ethics* 19: 449–465. https://doi.org/10.1177/17470161231180449
42. Hosseini, M., B. Gordijn, G. E. Kaebnick, and K. Holmes. 2025. Disclosing generative AI use for writing assistance should be voluntary. *Research Ethics*. https://doi.org/10.1177/17470161251345499
43. International Committee of Medical Journal Editors (ICMJE). 2024. *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals*. https://www.icmje.org/recommendations/
44. Jobin, A., M. Ienca, and E. Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1: 389–399. https://doi.org/10.1038/s42256-019-0088-2
45. Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589. https://doi.org/10.1038/s41586-021-03819-2
46. Kaebnick, G. E., D. C. Magnus, A. Kao, M. Hosseini, D. Resnik, V. Dubljević, et al. 2023. Editors' statement on the responsible use of generative AI technologies in scholarly journal publishing. *Hastings Center Report* 53: 3–6. https://doi.org/10.1002/hast.1507
47. Kapoor, S., and A. Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4: 100804. https://doi.org/10.1016/j.patter.2023.100804
48. Kasneci, E., K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103: 102274. https://doi.org/10.1016/j.lindif.2023.102274
49. Kendall, G., and J. A. Teixeira da Silva. 2024. Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learned Publishing* 37: 55–62. https://doi.org/10.1002/leap.1578
50. Kobak, D., R. González-Márquez, E.-Á. Horvát, and J. Lause. 2025. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances* 11: eadt3813. https://doi.org/10.1126/sciadv.adt3813
51. Latona, G. R., M. H. Ribeiro, T. R. Davidson, V. Veselovsky, and R. West. 2025. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.1145/3757667
52. Liang, W., M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4: 100779. https://doi.org/10.1016/j.patter.2023.100779
53. Liang, W., Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, et al. 2024a. *Mapping the increasing use of LLMs in scientific papers* [Preprint]. arXiv:2404.01268. https://doi.org/10.48550/arXiv.2404.01268
54. Liang, W., Z. Izzo, Y. Zhang, H. Lepp, H. Cao, X. Zhao, et al. 2024b. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the*

41st International Conference on Machine Learning (PMLR 235). https://proceedings.mlr.press/v235/liang24b.html

55. Liang, W., N. Rajani, X. Yang, E. Ozoani, E. Wu, Y. Chen, D. S. Smith, and J. Zou. 2024c. *What's documented in AI? Systematic analysis of 32K AI model cards* [Preprint]. arXiv:2402.05160. https://doi.org/10.48550/arXiv.2402.05160
56. Liu, X., S. Cruz Rivera, D. Moher, M. J. Calvert, A. K. Denniston, and the SPIRIT-AI and CONSORT-AI Working Group. 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine* 26: 1364–1374. https://doi.org/10.1038/s41591-020-1034-x
57. Liu, D., P. Nanayakkara, S. A. Sakha, G. Abuhamad, S. L. Blodgett, N. Diakopoulos, J. R. Hullman, and T. Eliassi-Rad. 2022. Examining responsibility and deliberation in AI impact statements and ethics reviews. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3514094.3534155
58. Lund, B. D., T. Wang, N. R. Mannuru, B. Nie, S. Shimray, and Z. Wang. 2023. ChatGPT and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology* 74: 570–581. https://doi.org/10.1002/asi.24750
59. Lund, B. D., and K. T. Naheem. 2024. Can ChatGPT be an author? A study of artificial intelligence authorship policies in top academic journals. *Learned Publishing* 37: 13–21. https://doi.org/10.1002/leap.1582
60. Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54: 115. https://doi.org/10.1145/3457607
61. Messeri, L., and M. J. Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature* 627: 49–58. https://doi.org/10.1038/s41586-024-07146-0
62. Metcalf, J., and K. Crawford. 2016. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society* 3. https://doi.org/10.1177/2053951716650211
63. Mitchell, M., S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, pp. 220–229. https://doi.org/10.1145/3287560.3287596
64. Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1: 501–507. https://doi.org/10.1038/s42256-019-0114-4
65. Mollaki, V. 2024. Death of a reviewer or death of peer review integrity? The challenges of using AI tools in peer reviewing and the need to go beyond publishing policies. *Research Ethics*. https://doi.org/10.1177/17470161231224552
66. Morley, J., L. Floridi, L. Kinsey, and A. Elhalal. 2020. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics* 26: 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

67. Munn, L. 2023. The uselessness of AI ethics. *AI and Ethics* 3: 869–877. https://doi.org/10.1007/s43681-022-00209-w

68. Nanayakkara, P., J. Hullman, and N. Diakopoulos. 2021. Unpacking the expressed consequences of AI research in broader impact statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3461702.3462608

69. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. U.S. Department of Health, Education, and Welfare.

70. National Institute of Standards and Technology (NIST). 2023. *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

71. Nature. 2023. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use [Editorial]. *Nature* 613: 612. https://doi.org/10.1038/d41586-023-00191-1

72. Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, et al. 2015. Promoting an open research culture. *Science* 348: 1422–1425. https://doi.org/10.1126/science.aab2374

73. Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366: 447–453. https://doi.org/10.1126/science.aax2342

74. Organisation for Economic Co-operation and Development (OECD). 2019. *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). OECD Publishing.

75. Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 372: n71. https://doi.org/10.1136/bmj.n71

76. Pushkarna, M., A. Zaldivar, and O. Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533231

77. Raji, I. D., A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, pp. 33–44. https://doi.org/10.1145/3351095.3372873

78. Resnik, D. B., and M. Hosseini. 2025a. Disclosing artificial intelligence use in scientific research and publication: When should disclosure be mandatory, optional, or unnecessary? *Accountability in Research*. https://doi.org/10.1080/08989621.2025.2481949

79. Resnik, D. B., and M. Hosseini. 2025b. The ethics of using artificial intelligence in scientific research: New guidance needed for a new tool. *AI and Ethics* 5: 1499–1521. https://doi.org/10.1007/s43681-024-00493-8

80. Rocher, L., J. M. Hendrickx, and Y.-A. de Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications* 10: 3069. https://doi.org/10.1038/s41467-019-10933-3

81. Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1: 206–215. https://doi.org/10.1038/s42256-019-0048-x

82. Secretary's Advisory Committee on Human Research Protections (SACHRP). 2022. *Considerations for IRB review of research involving artificial intelligence* (Attachment E, 25 July 2022). U.S. Department of Health and Human Services.

83. Selbst, A. D., d. boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, pp. 59–68. https://doi.org/10.1145/3287560.3287598

84. Shumailov, I., Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. 2024. AI models collapse when trained on recursively generated data. *Nature* 631: 755–759. https://doi.org/10.1038/s41586-024-07566-y

85. Stilgoe, J., R. Owen, and P. Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy* 42: 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

86. Stokel-Walker, C. 2023. ChatGPT listed as author on research papers: Many scientists disapprove. *Nature* 613: 620–621. https://doi.org/10.1038/d41586-023-00107-z

87. Strubell, E., A. Ganesh, and A. McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 3645–3650. https://doi.org/10.18653/v1/P19-1355

88. Thomas, J., and A. Harden. 2008. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology* 8: 45. https://doi.org/10.1186/1471-2288-8-45

89. Thorp, H. H. 2023. ChatGPT is fun, but not an author. *Science* 379: 313. https://doi.org/10.1126/science.adg7879

90. UNESCO. 2021. *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific and Cultural Organization.

91. van Dis, E. A. M., J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting. 2023. ChatGPT: Five priorities for research. *Nature* 614: 224–226. https://doi.org/10.1038/d41586-023-00288-7

92. Van Noorden, R., and J. M. Perkel. 2023. AI and science: What 1,600 researchers think. *Nature* 621: 672–675. https://doi.org/10.1038/d41586-023-02980-0

93. Walters, W. H., and E. I. Wilder. 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports* 13: 14045. https://doi.org/10.1038/s41598-023-41032-5

94. Wang, H., T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature* 620: 47–60. https://doi.org/10.1038/s41586-023-06221-2

95. Weber-Wulff, D., A. Anohina-Naumeca, S. Bjelobaba, T. Foltýnek, J. Guerrero-Dib, O. Popoola, et al. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity* 19: 26. https://doi.org/10.1007/s40979-023-00146-z

96. Weidinger, L., J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, pp. 214–229. https://doi.org/10.1145/3531146.3533088
97. Whittlestone, J., R. Nyrup, A. Alexandrova, and S. Cave. 2019. The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, pp. 195–200. https://doi.org/10.1145/3306618.3314289
98. Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18
99. Zielinski, C., M. A. Winker, R. Aggarwal, L. E. Ferris, M. Heinemann, J. F. Lapeña, et al. 2023. *Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications*. World Association of Medical Editors. https://wame.org/page3.php?id=106
100. Zook, M., S. Barocas, d. boyd, K. Crawford, E. Keller, S. P. Gangadharan, et al. 2017. Ten simple rules for responsible big data research. *PLOS Computational Biology* 13: e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Figures

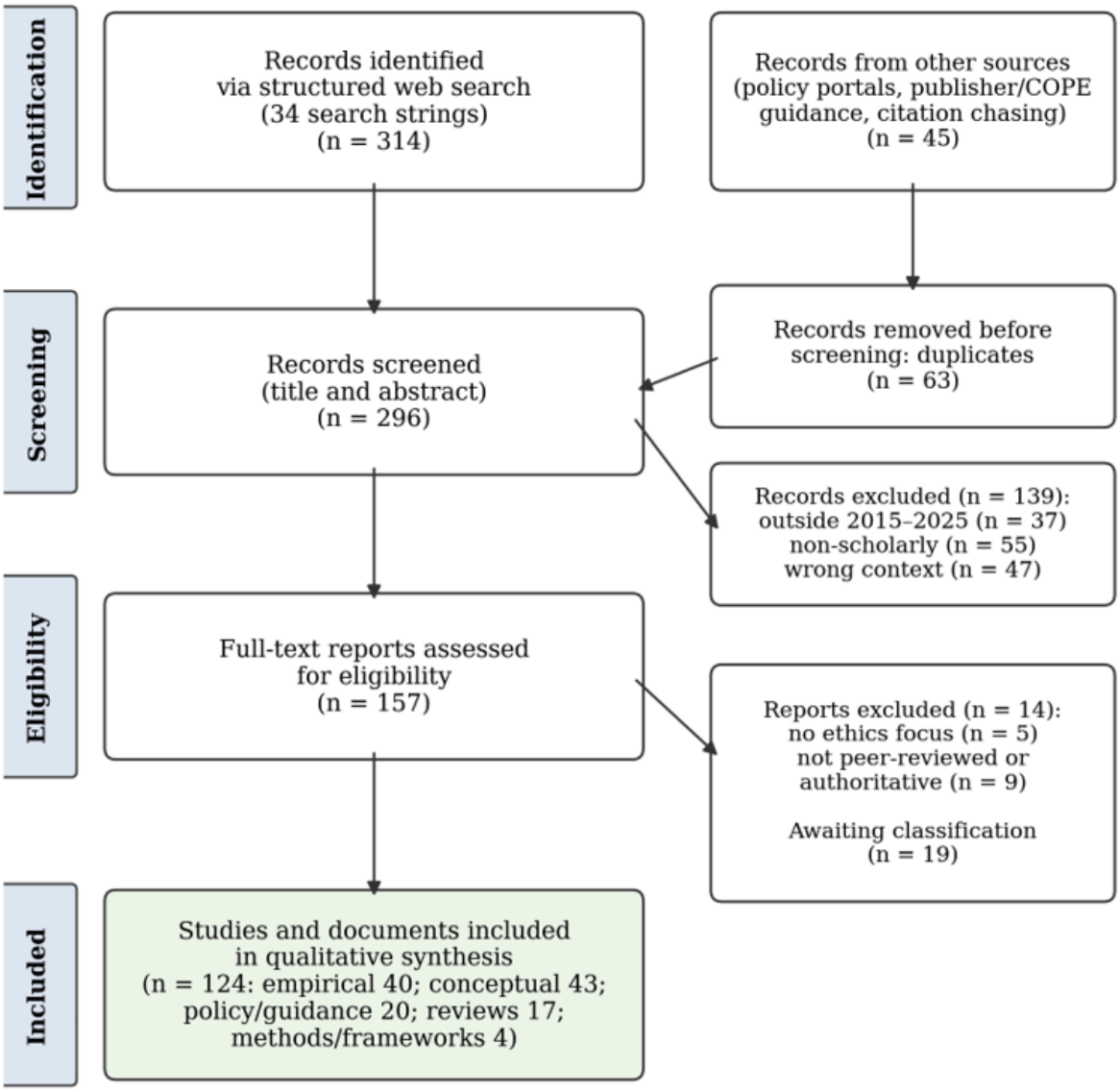


Figure 1

PRISMA 2020 flow diagram of study identification, screening, and inclusion (searches run 14 October 2026; publication window 2015–2025).

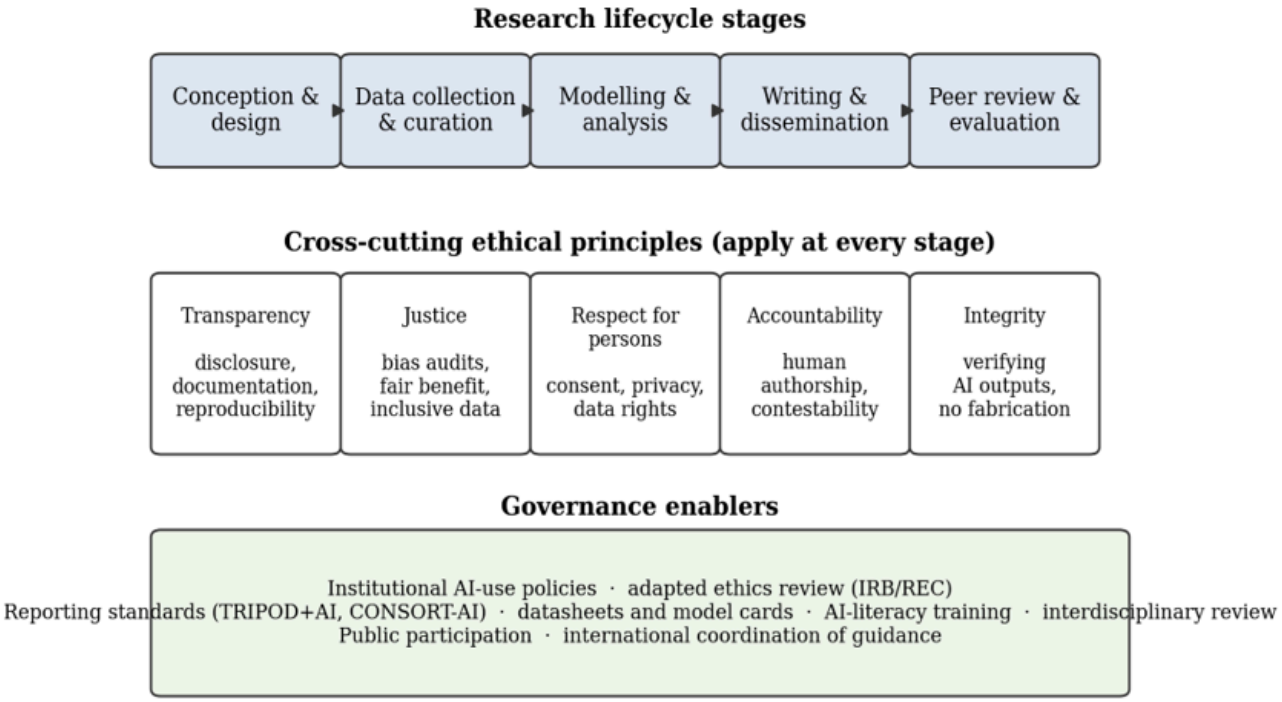


Figure 2

Integrated framework for ethical AI-enabled academic research, linking cross-cutting principles to research lifecycle stages and governance enablers (authors' synthesis).

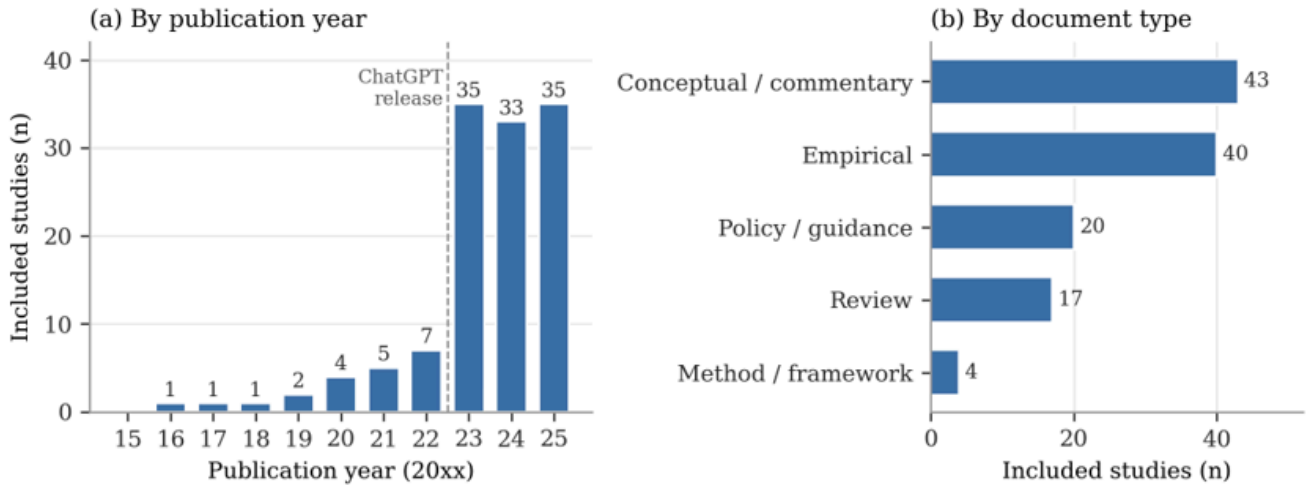


Figure 3

Included studies (n = 124) by (a) publication year and (b) document type. The dashed line marks the public release of ChatGPT (November 2022).