ARTICLE

MIT Science Policy Review

From policing authors to stewarding science: Governing artificial intelligence across the scientific publishing pipeline in medical journals

Rawan Abulibdeh,[1,2,3] Janan Arslan,[4,5] Sebastián Andrés Cajas Ordóñez,[6] Leo Anthony Celi,[7,8,9,*] Torleif Markussen Lunde,[10,11] and Mena Ramos[12]

Authors are listed in alphabetical order by last name. All authors contributed equally to the best of their abilities.

Edited by Sana Pandey and Austin Saragih

HIGHLIGHTS

- Most scientific journals have adopted AI policies, but governance remains narrowly focused on authorship and disclosure.
- Critical gaps persist in peer review, data integrity, enforcement, and equity—where AI increasingly shapes scientific claims.
- Comparative analysis reveals large variation across journals and regions, with many relying on umbrella guidance or lacking explicit policies.
- We introduce AI governance readiness levels (AGRL) to assess institutional preparedness for AI-mediated science.
- Using PRAIDE as an integrative example, we argue for community-owned, reflexive stewardship of the scientific knowledge system.

[1]TRANSFORM HF, Department of Cardiology, University Health Network, Toronto, ON, Canada
[2]Peter Munk Cardiac Centre, University Health Network, Toronto, ON, Canada
[3]Department of Electrical and Computer Engineering, University of Toronto, Ontario, Canada
[4]Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013 Paris, France
[5]Centre for Eye Research Australia, University of Melbourne, Royal Victorian Eye and Ear Hospital, East Melbourne, VIC 3002, Australia
[6]MIT Critical Data, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America
[7]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America
[8]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America
[9]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America
[10]Vestlandets Innovasjonsselskap, Postboks, Bergen, Norway
[11]University of Bergen, Bergen, Norway
[12]Global Ultrasound Institute, University of California San Francisco, San Francisco, California, United States of America
*Email: lceli@mit.edu

The authors declare no conflict of interest.

Generative artificial intelligence (AI) is rapidly transforming how scientific knowledge is produced, reviewed, and disseminated. In response, journals and publishing organizations have begun issuing policies to govern AI use in scholarly publishing. However, it remains unclear whether existing governance frameworks meaningfully address the risks AI introduces across the full publication pipeline. We conducted a narrative review of journal policies, publisher guidance, and recent analyses of AI governance in scientific publishing, complemented by direct examination of submission guidelines from high-impact, open-access, and regional medical journals. Our findings show that while most journals have converged on a narrow legal consensus (prohibiting AI authorship and requiring disclosure), yet governance remains fragmented and incomplete. Policies disproportionately target text generation by authors, while leaving critical domains under-regulated, including AI-assisted data analysis, peer review practices, enforcement mechanisms, and equity implications for researchers and reviewers globally. To synthesize these findings, we introduce the AI governance readiness levels, a five-level framework for assessing how well-equipped journals are to govern AI across the research and publication process. We further describe PRAIDE (Preparation, Representation, Attribution, Integrity checks, Dissemination, and Evaluation) as an illustrative architecture that integrates existing policies, integrity safeguards, and post-publication oversight into a coherent governance model. We argue that effective AI governance in scientific publishing cannot be achieved through static rules or journal-centric control alone. Instead, it requires a shift toward shared, adaptive oversight of the scientific publishing system, recognizing that responsibility for governing AI in science is collective, continuous, and inseparable from the public trust in research.

ARTIFICIAL intelligence is no longer a peripheral tool in scientific publishing. Large language models and related systems are increasingly used to draft manuscripts, analyze data, generate figures, assist peer reviewers, and support editorial decision-making [1] (a bibliometric analysis of the top 100 largest publishers and top 100 highest H-index journals on Scimago). This rapid integration has prompted journals, publishers, and professional organizations to issue guidance on acceptable AI use [2]. Yet these responses have emerged unevenly, often under time pressure, and with limited

coordination across the scientific ecosystem [3].

This paper argues that current AI governance in scientific publishing resembles a defensive strategy focused on the most visible risk, text generation and authorship, while leaving deeper structural vulnerabilities largely unaddressed [1, 3]. As AI systems increasingly influence how evidence is generated, evaluated, and legitimized, governance must extend beyond disclosure checklists toward mechanisms that safeguard research integrity, equity, and accountability across the full publication lifecycle. By examining existing policies and their blind spots, we aim to clarify what is known, what is missing, and what it would mean to move from policing AI use toward stewarding algorithmically mediated science.

In this paper, we make three primary contributions. First, we conduct a narrative review and policy landscape analysis of 11 global medical journals and two umbrella organizations, characterizing the distribution of governance attention across the full publication pipeline. Second, we introduce the AI governance readiness levels (AGRL), a five-level diagnostic framework for assessing the institutional capacity of journals to govern AI-mediated risks across the full publication pipeline. Finally, we propose collective, reflexive stewardship as the necessary reorientation for AI governance in scientific publishing, arguing that no single journal-centric policy approach is sufficient to address the systemic risks introduced by generative AI.

The current landscape of AI governance in scientific publishing

Methods: Literature review and policy collection.

We conducted a narrative review and policy landscape analysis to characterize how scientific journals and umbrella organizations are currently governing the use of generative artificial intelligence (AI) across the publishing pipeline, including authorship and writing support, figures and images, peer review, disclosure, and enforcement. A narrative approach was selected because the relevant evidence base spans heterogeneous sources (editorials, viewpoints, correspondence, empirical analyses, policy statements, and web-based "instructions for authors and reviewers") many of which are not indexed as academic publications. As a result, conventional systematic review methods are insufficient for comprehensive capture of journal-level AI governance.

Literature identification. We used a purposive search strategy designed to identify high-signal commentary and empirical analyses addressing AI use in scientific publishing and peer review. PubMed was searched on January 17, 2026, using three concept clusters: (i) generative AI and large language models (e.g., "ChatGPT," "generative AI," "large language model"); (ii) publishing governance and research integrity (e.g., "peer review," "editorial policy," "publication ethics," "research integrity"); and (iii) influential clinical journals that frequently function as policy bellwethers. Searches were limited to English-language publications from January 1, 2022 through December 31, 2025, corresponding to the period of rapid mainstream uptake of generative AI tools in academic writing and review. Records were screened for relevance to AI use in authorship, manuscript preparation, peer review, editorial processes, disclosure, and accountability, rather than clinical or methodological applications of AI. Database searching was complemented by forward and backward citation tracking and targeted inclusion of foundational policy and governance statements relevant to medical publishing, including guidance from international umbrella organizations. The PubMed search retrieved 14 records. Two reviewers (T.M.L. and S.A.C.O.) independently screened all 14 records by title and abstract. Records were included if they addressed AI use in authorship, manuscript preparation, peer review, editorial processes, disclosure, or accountability in scientific publishing. Records were excluded if they addressed AI exclusively in clinical care, data extraction methodology, or administrative coding without publishing implications, or if AI was mentioned only tangentially. Of the 14 records, 7 met inclusion criteria and were retained for synthesis; 7 were excluded (clinical AI applications without publishing implications, n=4; AI methods for systematic reviews without policy discussion, n=2; administrative data coding, n=1). Disagreements were resolved by consensus.

Journal and organization policy sampling. To assess what journals actually regulate, distinct from what the literature recommends, we assembled an exploratory purposive sample of journals intended to reflect the contemporary global clinical publishing ecosystem. Rather than sampling from a formally enumerated database frame, the research team identified candidate journals on the basis of domain knowledge of the clinical publishing ecosystem, targeting three functional roles within it. Journals were accordingly selected across three tiers: (i) high-impact general medical journals that function as policy "originators"; (ii) high-visibility open-access and digitally native journals that function as policy "diffusers"; and (iii) regional journals based in Africa, Asia, and Latin America, where governance capacity and policy discoverability may differ. Within each tier, individual journals were proposed by team members and cross-checked collectively before inclusion. The principal operative requirement for inclusion was that a journal publish author or reviewer instructions retrievable at the time of policy capture, since the analysis depended on direct examination of those guidelines; where AI-relevant guidance could not be located, this was recorded as an empirical finding rather than treated as a silent exclusion. In addition, two umbrella organizations were included because of their role in shaping downstream journal policy adoption and harmonization. The goal of this sampling strategy was not exhaustiveness or statistical representativeness, but structural coverage of influential policy originators, policy diffusers, and settings where governance capacity may differ. Although the sample is small relative to the global publishing ecosystem, in which AI policies have proliferated rapidly across hundreds of journals and publishers [2, 4], it was designed to capture the actors most likely to shape, absorb, and transmit governance norms. We acknowledge that this expert-driven, non-enumerative

strategy does not support claims of representativeness, a limitation we address below. This design nonetheless enabled analysis of both convergence in AI governance norms and structural asymmetries across regions and publishing models. The final sample comprised 11 journals (distributed across the three tiers) and two umbrella organizations.

Policy discovery and collection. Because journal AI policies are frequently embedded in web pages, author instructions, editorial guidance, publisher policy hubs, or intermittently updated statements, rather than in stable, indexed documents, we employed a multi-step, hand-search approach. For each journal and organization, we: (1) manually navigated journal websites to locate AI-related guidance within instructions for authors and reviewers, editorial policy pages, and publisher-level policy portals; (2) when policies were fragmented or difficult to locate, used targeted link-retrieval queries to identify relevant policy URLs; (3) saved local copies of key policy pages as PDFs (or archived versions when available) to mitigate instability from web updates; and (4) resolved discrepancies by returning to the primary source and prioritizing the most recent, explicitly versioned policy language where possible. This process was particularly important for regional journals, where site architecture, indexing, and policy consolidation were often inconsistent. In such cases, the absence or poor discoverability of AI governance statements was treated as an empirical governance finding, reflecting practical barriers to compliance and enforcement, rather than assumed to represent researcher oversight.

Policy extraction and verification. Policies were extracted using a structured template aligned with core checkpoints across the publication pipeline: AI authorship; AI-assisted writing; figures and images; peer review; disclosure requirements (including location and level of specificity); and enforcement or accountability mechanisms. Extraction followed a hybrid workflow: initial automated or LLM-assisted parsing to organize policy text into predefined categories, followed by manual verification against the primary source to confirm accuracy and capture qualifying language, exceptions, and cross-references to umbrella standards or publisher-level policies. Where journals deferred to external guidance rather than articulating journal-specific rules, these dependencies were recorded explicitly to distinguish "inherited" governance from "native" policy development.

Synthesis. Findings were synthesized narratively, with emphasis on: (i) areas of cross-journal convergence, such as the prohibition of AI authorship and the non-delegability of human accountability; (ii) domains of heterogeneity, particularly with respect to disclosure specificity and governance of peer review; and (iii) governance gaps revealed through absent, ambiguous, or difficult-to-locate policies. This landscape analysis is descriptive and diagnostic, forming the empirical foundation for subsequent examination of governance readiness and for the integrative PRAIDE framing introduced later in the manuscript.

What the literature converges on.

Across editorials, commentaries, empirical analyses, and policy statements published since the widespread uptake of generative AI tools, a limited but coherent set of normative convergences has emerged regarding their role in scientific publishing. While implementation details vary across journals and disciplines, the literature demonstrates consistent agreement on several foundational principles related to authorship, accountability, and transparency. At the same time, these areas of convergence are notably narrow, leaving substantial portions of the publication pipeline under-theorized and weakly governed.

AI cannot be an author and accountability remains non-delegable. There is consensus among major journals and umbrella organizations that generative AI systems cannot meet authorship criteria and should not be listed as authors on scientific manuscripts [1, 5, 6]. This position is articulated consistently by major journals, publishers, and umbrella organizations, including the International Committee of Medical Journal Editors (ICMJE) and the Committee on Publication Ethics (COPE), and has been rapidly adopted across the clinical publishing ecosystem [5].

The rationale for this prohibition is not primarily technological, but normative and legal. AI systems cannot assume responsibility for the integrity, originality, or ethical conduct of research; cannot approve a final manuscript; and cannot be held accountable for errors, conflicts of interest, or misconduct [5, 6]. As a result, the literature treats human accountability as non-delegable (meaning it cannot be transferred to an AI system or third party): responsibility for all submitted content remains with human authors, regardless of whether AI tools are used for drafting, editing, analysis, or revision [5, 7]. This principle is framed as foundational rather than negotiable, reflecting concern that delegating epistemic responsibility to automated systems would undermine core norms of scientific integrity [3].

Disclosure is necessary, but weakly standardized. A second point of convergence is the recognition that transparency about AI use is essential to maintaining trust in the scientific record. Across disciplines, authors emphasize that undisclosed AI assistance risks obscuring intellectual contributions, complicating accountability in cases of error or retraction, and introducing opaque sources of bias or fabrication [1, 3, 5]. As a result, disclosure of AI use is now widely endorsed as a baseline governance requirement.

However, while the need for disclosure is broadly accepted, the literature reveals substantial disagreement regarding how and where such disclosure should occur, and what level of detail is meaningful [1]. Proposed practices range from brief acknowledgments of AI-assisted writing to detailed reporting of tool names, versions, functions, and prompts [5, 7]. Recent consensus reporting guidelines, such as the GAMER Statement for generative AI use in medical research [8], seek to standardize these expectations through defined items covering tool specifications, prompting, the tool's role, content

verification, and data privacy. This lack of standardization creates enforcement challenges and prevents systematic tracking of AI usage in the scientific record [1, 3]. This tension recurs across journals and publication formats, suggesting that disclosure has become a shared norm without a shared operational standard.

Governance has focused narrowly on writing assistance. A recurring observation across the literature is the disproportionate focus on AI-assisted writing relative to other stages of the publication process [1, 9]. Most published guidance addresses text generation, language editing, and plagiarism-related concerns, while comparatively little attention is given to AI use in data analysis, statistical inference, figure or image generation, peer review, or editorial decision-making [1, 9, 10].

Several authors explicitly caution that this narrow framing risks mischaracterizing the primary epistemic risks (risks to the reliability and validity of scientific knowledge) introduced by generative AI [3, 9, 11]. Errors arising from AI-assisted data interpretation, fabricated or manipulated images, or automated review support may be less visible than text reuse, but potentially more consequential for scientific validity and downstream clinical decision-making [9, 11]. The literature thus converges on the view that current discussions of AI governance in publishing capture only a subset of the risks introduced by these tools.

Acknowledged mismatch between practice and policy. Finally, the literature consistently notes a growing mismatch between rapidly evolving AI-enabled research practices and comparatively static journal policies [1, 3]. Multiple analyses highlight that AI tools are already embedded in routine scholarly workflows, often preceding or outpacing formal policy development [1, 9]. A 2025 survey of 1,645 researchers found that 53% of reviewers had used AI during peer review, though only 19% for analytical tasks such as assessing methodology or statistics [12]. The gap between use and disclosure is stark: across 25,114 manuscripts submitted to 49 biomedical journals, only 5.7% disclosed any AI use, far below the 50–76% usage reported in surveys [13]. In the absence of clear, enforceable standards, authors, reviewers, and editors frequently rely on informal norms, inherited publisher guidance, or individual judgment [10].

Importantly, this mismatch is rarely framed as institutional negligence. Rather, it is understood as a structural challenge arising from the speed of technological change, decentralized governance, and the historical role of journals as reactive rather than anticipatory regulatory actors [3]. Several authors caution that reliance on ad hoc disclosures or checklist-style requirements may create a false sense of security without addressing deeper questions of accountability, oversight, and epistemic risk [1, 3, 11].

Across disciplines, the literature converges on a narrow set of shared commitments: AI systems are not authors; accountability remains irreducibly human; disclosure of AI use is ethically required; and existing governance frameworks are both fragmented and lagging. Beyond these minimal points of agreement, however, consensus quickly fractures, particularly around questions of scope, enforceability, and the legitimate boundaries of machine assistance in scholarly work. It is within this gap between declarative principle and operational practice that the central problem of AI governance in scientific publishing now resides, demanding closer empirical scrutiny of how journals translate ethical claims into editorial action.

What journals actually regulate.

To move beyond the normative register of the existing literature, we examined how claims about AI governance are translated into the everyday regulatory work of scientific publishing. This section presents empirical findings from a structured analysis of journal- and umbrella-organization policies, tracing what is formally governed, where obligations are situated across the publication pipeline, and how responsibility and enforcement are defined, or left conspicuously indeterminate.

Journal sample description. Using the three-tier sample described in the Methods, our policy analysis included 11 medical journals and two umbrella organizations, the International Committee of Medical Journal Editors (ICMJE) and the Committee on Publication Ethics (COPE), selected to capture variation in influence, publishing model, and geographic context. The umbrella organizations were included because their guidance is frequently referenced verbatim or implicitly adopted by downstream journals, shaping policy harmonization across publishers.

Table I summarizes the journal sample, including journal type, publisher affiliation, geographic orientation, and the presence or absence of explicit AI-related policy language. Regional inclusion was intentional rather than illustrative. Difficulty locating, interpreting, or verifying AI governance statements for some journals, particularly outside North America and Europe, was treated as a material feature of the governance landscape, rather than as a limitation of data collection.

Core policy dimensions across the publication pipeline. Across the sampled journals, AI governance was unevenly distributed across stages of the publication process. We identified six recurring policy dimensions: authorship; AI-assisted writing; figures and images; peer review; disclosure requirements; and enforcement or accountability mechanisms [1]. Table II presents a policy matrix mapping each journal against these dimensions.

Several patterns emerged. First, authorship and AI-assisted writing were the most consistently regulated domains. Most journals explicitly prohibited AI systems from being listed as authors and permitted limited use of AI for writing or editing under conditions of human responsibility [1, 6]. Second, governance attenuated sharply beyond manuscript preparation. Policies addressing AI use in peer review, editorial workflows, data analysis, or figure generation were sparse, ambiguous, or entirely absent in many journals [1, 10].

TABLE I: Journal sample characteristics included in the policy analysis.

Tier	Journal/Organization	Country/Region	Publisher/Affiliation	Notes on policy provenance
High-impact Western	New England Journal of Medicine (NEJM)	USA	Massachusetts Medical Society	Journal-level policy
	The Lancet	UK	Elsevier	Journal policy + publisher-level policy dependencies
	JAMA	USA	American Medical Association	Journal/network policy
Open access	PLOS Medicine/PLOS ONE	USA	Public Library of Science	Journal-level policy; explicit author/reviewer guidance
	Journal of Medical Internet Research (JMIR)	Canada	JMIR Publications	Journal-level policy; granular disclosure expectations
Regional: Africa	South African Medical Journal (SAMJ)	South Africa	Health & Medical Publishing Group (HMPG)	Journal-level policy page
	Pan African Medical Journal (PAMJ)	Kenya	Independent	Journal policy aligned with external standards (e.g., WAME/JAMA)
Regional: Asia	Chinese Medical Journal (CMJ)	China	Chinese Medical Association/Lippincott	Publisher-template influence noted
	Journal of Korean Medical Science (JKMS)	South Korea	Korean Academy of Medical Sciences (KAMS)	Journal-level policy; explicit enforcement language reported
Regional: Latin America	Clinics	Brazil	Hospital das Clínicas/USP	Reliance on umbrella guidance implied; limited explicit AI policy located
	Brazilian Journal of Medical and Biological Research (BJMBR)	Brazil	Associação Brasileira de Divulgação Científica (ABDC)	Reliance on umbrella guidance implied; limited explicit AI policy located
Umbrella organizations	International Committee of Medical Journal Editors (ICMJE)	International	ICMJE	Foundational recommendations; widely referenced
	Committee on Publication Ethics (COPE)	International	COPE	Position statement; widely referenced

Note: For some journals, difficulty locating explicit AI policy language (or reliance on umbrella guidance without local specification) was treated as a governance-relevant empirical finding rather than a data-collection failure.

TABLE II: AI governance dimensions across the journal sample (policy comparison matrix).

Journal/Org	AI as author	Writing assist.	Figures/images	Peer review	Disclosure	Enforcement
NEJM	P	A	A	R	Y	N/E
The Lancet	P	R	P	A	Y	Y
JAMA	P	A	D	P	Y	Y
PLOS (Med/ONE)	P	A	A[a]	P	Y	Y
JMIR	P	A	A	A[b]	Y	Y
SAMJ	P	A	A	A	Y	N
PAMJ	P	A	A	P	Y	Y
CMJ	P	A	A	N/S	Y	N
JKMS	P	A	A	N/S	Y	Y
Clinics	I	I	N/S	N/S	I	N
BJMBR	I	I	N/S	N/S	I	N
ICMJE	P	A	A	N/S	Y	N/A
COPE	P	A	A	A	Y	N/A

Codes: AI as author: P=prohibited; I=implicitly prohibited via ICMJE-aligned authorship criteria. Writing assistance/Figures/Peer review: A=allowed (typically with conditions); R=restricted (limited use cases); D=discouraged; P=prohibited; N/S=not specified. Disclosure: Y=explicitly required; I=implied (e.g., via umbrella guidance without journal-specific detail). Enforcement: Y=explicit consequences/process named; N=no explicit enforcement mechanism; N/E=not explicit/unclear; N/A=not applicable.
Interpretation rule: "Not specified" denotes no explicit, retrievable AI-specific policy language for that dimension at time of policy capture, and is treated as a descriptive result.
[a] PLOS policies often distinguish permissible image enhancement from prohibited fabrication/manipulation.
[b] JMIR permits reviewer AI use with restrictions (including confidentiality/terms-of-service constraints and disclosure).

A substantial proportion of journals deferred to external standards (most commonly ICMJE or COPE) rather than articulating journal-specific rules [1, 14]. While such reliance may promote harmonization, it also obscures the locus of operational responsibility and introduces variability in how policies are interpreted and applied at the journal level.

Disclosure requirements. Disclosure of AI use was the most frequently cited governance mechanism across journals, but requirements varied markedly in location, specificity, and scope.

Some journals required disclosure only when AI tools contributed substantively to manuscript drafting, while others mandated disclosure for any AI-assisted activity, including editing or formatting. Disclosure locations ranged from acknowledgments sections to methods sections, cover letters, or submission checklists [1, 5]. Few journals specified how disclosures should be evaluated by editors or reviewers, or how disclosed information should inform editorial decision-making.

High-impact and digitally native journals were more likely to request granular disclosures, such as tool names, versions, and descriptions of functional use [7]. In contrast, many journals relied on high-level language (e.g., "authors should disclose use of AI tools") without defining thresholds, formats, or verification mechanisms [1]. Regional journals, in particular, often adopted generalized disclosure language or deferred entirely to publisher-level or umbrella guidance.

This heterogeneity complicates cross-journal comparability and creates uncertainty for authors submitting to multiple venues, particularly when disclosure expectations are implicit, fragmented, or distributed across multiple policy documents [1].

Enforcement & accountability. Explicit enforcement mechanisms were inconsistently articulated across the journal sample. While many policies affirmed that authors retain full responsibility for submitted content [5], few specified consequences for non-disclosure or misuse of AI tools. Some journals, such as the Journal of Korean Medical Science, established relatively explicit enforcement positions, and editorial commentary from other venues (e.g., BMJ [15]) has described comparably explicit positions, but many others did not provide explicit enforcement pathways [1].

Policy language was frequently aspirational rather than operational, emphasizing ethical conduct and transparency without detailing enforcement pathways. Peer review emerged as the least governed domain, with limited guidance on whether reviewers may use AI tools, how such use should be disclosed, or how confidentiality, bias, and data handling concerns should be addressed. Where enforcement was mentioned, it varied in specificity and integration with broader research integrity frameworks [1, 16].

These findings show that journal-level AI governance remains fragmented, uneven in scope, and disproportionately concentrated on questions of manuscript authorship and writing. Reliance on umbrella organizations is widespread, yet this delegation does not resolve the absence of operational specificity at the level where editorial decisions are enacted. Crucially, the absence, ambiguity, or limited discoverability of policies—particularly in regional and non-dominant publishing contexts, should be understood not as a neutral gap but as an active governance condition, with direct consequences for accountability, participation, and epistemic equity.

PRAIDE as an integrative synthesis.

The empirical findings described above point not merely to heterogeneity across journal AI policies, but to a deeper structural fragmentation in how governance is conceptualized within scientific publishing. Policy responses are predominantly reactive, anchored to isolated moments in the publication process, and heavily reliant on inherited or umbrella guidance [1]. As a result, governance is dispersed across documents, actors, and timepoints, with limited visibility into how obligations, risks, and downstream effects accumulate across the full lifecycle of a scientific article.

This fragmentation becomes more legible when situated against established models of peer review. Traditional publishing workflows have long conceptualized peer review as a bounded sequence—editorial triage, reviewer assessment, iterative revision, and publication decision-making—within which integrity safeguards are expected to operate. These models provided a relatively stable baseline prior to the widespread integration of generative AI [17]. Increasingly, however, empirical evidence from policy analyses, surveys, and documented integrity failures suggests that AI now shapes research design, analysis, writing, and evaluation well before and well beyond these formal checkpoints [1]. Governance frameworks have struggled to adapt to this expanded temporal and functional footprint, producing a persistent gap between stated ethical commitments and observed practice.

To address this disjunction, we introduce PRAIDE as an integrative analytic framework for situating AI governance across the full peer review and publishing process. PRAIDE was developed in response to a core question emerging from the literature and policy review: *where, within the contemporary publication workflow, must integrity checkpoints exist in order to meaningfully address AI-related risks?* The framework maps governance considerations across six interrelated stages - Preparation, Representation, Attribution, Integrity checks, Dissemination, and Evaluation - spanning submission through post-publication oversight. By shifting attention from individual rules or policy documents to the workflow as a whole, PRAIDE makes visible where governance is explicit, implicit, deferred, or absent. Figure 1 presents PRAIDE as a visual synthesis, illustrating how AI governance considerations may be distributed across stages of scientific publishing without implying a single optimal or prescriptive model.

PRAIDE is grounded in three converging insights from the empirical evidence. First, while professional bodies and leading journals have articulated clear norms around authorship and human accountability [6], these norms alone are insufficient when AI tools are embedded within routine research and review practices [1]. Second, a persistent policy–practice gap remains evident: AI use by authors and reviewers is widespread, yet disclosure requirements and oversight mechanisms are inconsistently applied [1]. Third, documented failures of integrity safeguards—including cases in which statistical anomalies, data provenance issues, or methodological inconsistencies were not identified during

PRAIDE: Preparation, Representation, Attribution, Integrity Checks, Dissemination, and Evaluation

Enhanced Peer Review Workflow for Medical Journals in the AI Era

Incorporating AI Policies, Statistical Integrity Checks, Data Verification & Author Accountability Measures

Baseline framework adapted from Drozdz & Ladomery (2024) "The Peer Review Process: Past, Present, and Future"

Original Process | AI Policy Steps | Data Integrity Safeguards | Statistical / Plausibility Checks | Author Accountability | Community Engagement

Case Study: Image Manipulation by AI in the New England Journal of Medicine (2026)

An "Images in Clinical Medicine" item in the *New England Journal of Medicine* (April 2026; Wang & Mu) showed black airway "casts" from an 87-year-old man after forest-fire smoke exposure. The authors acknowledged using an AI tool to reposition the measuring ruler, and the item was retracted after a reader flagged the doctored scale on PubPeer. It was the journal's first retraction since 2020, when NEJM pulled a COVID-19 paper that relied on data from the company Surgisphere.

Image Manipulation Red Flags:

- Scale bar beginning at an implausible value
- Ruler digits that do not count sequentially
- Suspiciously uniform, unusually robust structures
- AI edits invisible to reviewers

Governance & Oversight Gaps:

- AI image edit not disclosed at submission
- Authors unaware of the journal's image policy
- Alteration cleared peer review at a top journal
- Caught only by scrutiny after publication

Outcome: *Retracted at the authors' request for an undisclosed AI image edit. If a scale bar altered with AI can clear review at medicine's most selective journal, the traceability of every figure is in question.*

Submission & Pre-Review Screening (Steps 1–8)

1. Manuscript submitted to journal
Ref: Drozdz & Ladomery (2024)

2. Author Confirmation & Accountability Waiver
ALL authors individually sign/verify: (1) Meet ICMJE 4 criteria; (2) Approve final manuscript; (3) Confirm data integrity & provenance; (4) Accept accountability for all aspects.
Ref: ICMJE Recommendations; Elsevier Co-Author Verification; JMIR License to Publish; Ochsner Journal Attestation Form

3. AI Use Disclosure Required
Declare AI tools, versions, prompts & methodology; AI cannot be listed as author; humans accountable for accuracy.
Ref: ICMJE; GAMER Statement (2025); Macdonald & Abbasi, BMJ (2023); Ganjavi et al., BMJ (2024); AlFayyad et al. (2025): only 5.7% disclosed AI use vs 50–76% actual

4. Plausibility & Scope Assessment
Author experience match study scope? · Verify data access (IRB, agreements) · Timeline vs. affiliation · Sample size vs. claimed precision? · Participant numbers verifiable?
Ref: COPE Guidelines; Bagenal, Lancet (2024); Frontiers (2025)

5. Data Verification & Provenance Check
Clinical/Real-World: IRB/Ethics docs; Trial registration (ClinicalTrials.gov); CDISC/CDASH compliance; EDC audit trails (21 CFR Part 11). *Synthetic/Augmented:* Validation framework; fidelity & privacy metrics; generation methodology transparent.
Ref: CDISC Standards; Azizi et al., BMJ Open (2021); Chen et al., BMC (2019); Della Porta et al. (2023)

6. Non-qualitative editorial review — Rejected ← if fails
Ref: Drozdz & Ladomery (2024)

7. Reference Quality & Completeness Check
Complete pub. info · Check for retracted papers · Verify citations support claims · Flag excessive self-citation patterns.
Ref: COPE Retraction Guidelines; Retraction Watch Database

8. Image Integrity & Forensic Screening
AI-powered detection of duplicated/manipulated figures; cross-reference against publication image databases.
Ref: ORI Image Forensics Guidelines; Bik et al., mBio (2016)

Editorial Screening & Peer Review (Steps 9–19)

9. Statistical Anomaly Detection
All $p<0.001$ with no null findings · Uniform effect sizes across diverse outcomes · Implausible claims · "Too clean" data patterns inconsistent with real-world variability.
Ref: Carlisle, Anaesthesia (2017): GRIM/SPRITE; Head et al., PLOS Biol (2015)

10. AI Detection & Compliance Screening
Plagiarism check + AI-generated content detection + disclosure verification.
Ref: Frontiers (2025): 53% of reviewers use AI; only 19% for analytical tasks; 71% concerned

11. Qualitative editorial review
Ref: Drozdz & Ladomery (2024)

12. Reviewers selected; method chosen
Ref: Drozdz & Ladomery (2024)

13. Mandatory Cross-Disciplinary Review
Interdisciplinary papers require reviewers from EACH relevant field; address potential affiliation bias.
Ref: von Wedel et al., JAMA (2024): LLM affiliation bias in peer review; Frontiers (2025)

14. Reviewer AI Policy Acknowledgment
Reviewers confirm: No uploading manuscripts to commercial AI (confidentiality); disclose AI assistance; human judgment required.
Ref: Perlis et al., JAMA (2025); Macdonald & Abbasi, BMJ (2023): "breach of confidentiality"

15. Reviewer Access to Underlying Data
Reviewers can request access to raw data, analysis code, protocols, and statistical outputs.
Ref: ICMJE Data Sharing Statement requirements; CONSORT 2010; STROBE Guidelines

16. AI-Assisted Quality & Bias Assessment
Use validated tools with human oversight; AI assists but humans make final determinations.
Ref: PROBAST+AI (Moons et al., BMJ 2025); RobotReviewer (Arno et al., Ann Intern Med 2022); Frontiers (2025)

17. Review carried out; reports returned
Ref: Drozdz & Ladomery (2024)

18. AI Content Verification Check
Verify AI-generated content accuracy; confirm disclosures complete; check for AI "hallucinations."
Ref: Gartlehner et al., Ann Intern Med (2025): AI data extraction error rates; Frontiers (2025)

19. Editorial decision
Ref: Drozdz & Ladomery (2024)
Accepted | Minor revisions | Major revisions | Rejected

Decision & Post-Publication Oversight (Steps 20–27)

20. Final Author Re-Confirmation (Pre-Publication)
All authors re-verify approval of final version; confirm no changes to authorship without explanation.
Ref: ICMJE: "editors should seek explanation and signed statement for requested change from all listed authors"

21. Publication
Ref: Drozdz & Ladomery (2024)

22. Post-Publication AI Monitoring
Ongoing monitoring for undisclosed AI use; community flagging; citation network analysis.
Ref: Frontiers (2025): "post-publication surveillance"; COPE Post-Publication Discussions Guidelines

23. Live Discussion & Community Dialogue
Real-time commenting/annotation on published papers; immediate flag response; open scholarly discourse.
Ref: Frontiers (2025) Community Pillar; PubPeer; Hypothes.is; F1000Research open review model

24. Protected Whistleblower Pathway
Anonymous reporting mechanism; protection for reporters; rapid response protocol.
Ref: ORI Whistleblower Protections; COPE Whistleblowing Guidelines

25. External Expert Challenge Pathway
Clear channel for domain experts to challenge published claims; rapid institutional response mechanism.
Ref: Case lesson: AI-altered image surfaced via post-publication scrutiny (PubPeer); COPE Responding to Concerns Guidelines

26. Cross-Institutional Fraud Alert Network
Shared database of concerns; coordinated investigations; track citation spread of problematic papers.
Ref: ORI Case Summaries Database; Retraction Watch Database; PubPeer

27. Correction / Retraction Pathway
Clear labeling; linked retractions; citation alerts to authors who cited problematic work.
Ref: COPE Retraction Guidelines; ICMJE Corrections/Retractions Requirements

AI Policy Sources
ICMJE Recommendations · GAMER Statement (2025) · Macdonald & Abbasi, BMJ (2023) · Perlis et al., JAMA (2025) · Ganjavi et al., BMJ (2024) · Frontiers AI Report (2025)

Statistical Integrity Sources
Carlisle, Anaesthesia (2017) · Head et al., PLOS Biol (2015) · PROBAST+AI (Moons 2025) · von Wedel et al., JAMA (2024) · Arno et al., Ann Intern Med (2022)

Data Integrity Sources
COPE Guidelines · ORI Resources · CDISC/CDASH Standards · 21 CFR Part 11 · CONSORT 2010 · STROBE · Retraction Watch Database

Author / Reviewer Accountability
ICMJE 4 Criteria · CRediT Taxonomy · Elsevier Co-Author Verification · JMIR License to Publish · Springer Nature Identity Verification

Frontiers AI Governance Framework (2025): Six Pillars for Responsible AI Integration

Key findings: 53% of reviewers use AI; only 19% for analytical tasks; 35% self-taught; 71% concerned about misuse; only 21% trust AI in publishing

1. Transparency	2. AI Literacy	3. Ethics Guardrails	4. Equity & Access	5. Community	6. Policy Leadership
Disclose when/how AI is used	Structured training programs	Human accountability required	Reduce systemic bias	Feedback mechanisms	Evidence-based standards

Complete Reference List

Drozdz A, Ladomery M. (2024) *Front Cell Dev Biol* · Arno A et al. (2022) *Ann Intern Med* · von Wedel D et al. (2024) *JAMA* · Perlis RH et al. (2025) *JAMA* · Gartlehner G et al. (2025) *Ann Intern Med* · Macdonald H, Abbasi K. (2023) *BMJ* · Ganjavi C et al. (2024) *BMJ* · Moons KGM et al. (2025) *BMJ* · Wang Y, Mu X. (2026) *N Engl J Med* (retracted) · AlFayyad I et al. (2025) *medRxiv*

Bagenal J. (2024) *Lancet* · Frontiers Media (2025) *Unlocking AI's Untapped Potential* · Carlisle JB. (2017) *Anaesthesia* · Head ML et al. (2015) *PLOS Biol* · Bik EM et al. (2016) *mBio* · Azizi Z et al. (2021) *BMJ Open* · Chen J et al. (2019) *BMC Med Inform* · Luo X et al. (2025) *BMJ Evid Based Med* (GAMER) · ICMJE, COPE, WAME, ORI Guidelines

Figure 1: The PRAIDE framework as an integrative synthesis of AI governance across the scientific publishing lifecycle. PRAIDE maps governance checkpoints across six stages: Preparation, Representation, Attribution, Integrity checks, Dissemination, and Evaluation, spanning submission through post-publication. The framework is descriptive rather than prescriptive, enabling comparison of where existing policies concentrate, defer, or omit AI governance across the publication pipeline.

peer review—highlight the limits of governance mechanisms that operate only at discrete or downstream points in the publication process.

By aligning governance expectations to stages of the workflow rather than to specific actors or tools, PRAIDE provides a common analytic structure for comparing how journals and organizations address similar risks using different terminologies, policy locations, and enforcement strategies. The framework does not assume harmonization or uniform institutional capacity. Instead, it enables heterogeneous and unevenly scoped policies to be examined coherently, revealing systematic clustering of governance attention around certain stages—most notably attribution and disclosure—and relative under-specification elsewhere, including peer review practices, data integrity, and post-publication evaluation.

Importantly, PRAIDE is descriptive rather than prescriptive. It does not define minimum standards, assign responsibilities to particular actors, or specify enforcement thresholds. Nor does it presume that all stages of the publication lifecycle require equivalent levels of regulation. Its purpose is analytic: to clarify how existing policies collectively shape the integrity of the scientific record, and where accountability becomes diffuse or displaced as AI-related risks propagate across the publication pipeline. PRAIDE also augments an existing baseline workflow [17] rather than constituting a validated protocol: it does not specify the time, personnel, or cost associated with individual checkpoints, and the steps it depicts are distributed across authors, editors, automated screening, reviewers, and post-publication oversight rather than added wholesale to any single reviewer's workload.

Within this analysis, PRAIDE functions as a synthesis tool linking the empirical policy landscape to questions of AI governance readiness. By rendering visible the misalignment between where risks emerge and where governance currently concentrates, the framework provides a structured basis for assessing variation in journals' capacity to operationalize AI oversight.

Gaps, blind spots, and AI governance readiness

Structural gaps in existing policies. The landscape analysis suggests that contemporary AI governance in scientific publishing is shaped less by isolated policy omissions than by a deeper structural mismatch between how AI is now used across research and review workflows and how journal oversight has historically been conceived and operationalized. Journals have moved rapidly to articulate prohibitions on AI authorship and to endorse disclosure as a normative principle [1, 5]. Yet the resulting policy architecture remains narrowly scoped, unevenly specified, and weakly supported by operational infrastructure. The gaps identified below reflect recurring structural patterns across the publishing ecosystem rather than failures of individual journals or actors.

The most visible and consistently governed domain remains authorship and AI-assisted writing, where the now-settled prohibition on AI authorship establishes an important normative floor but concentrates attention on a largely resolved question. By contrast, AI tools are increasingly embedded in practices that shape scientific claims more directly (data generation or extraction, statistical analysis, image production, and evidence synthesis), yet policies governing these uses remain sparse, uneven, or absent [1, 9]. The net effect is a governance regime optimized for regulating who wrote the manuscript rather than how evidence was produced, evaluated, and validated in AI-mediated workflows [1, 9].

A second structural gap concerns limited attention to data integrity and provenance in AI-mediated workflows. Even where journals permit or anticipate AI use in analytic tasks, few policies specify expectations for documenting data sources, audit trails, or provenance when AI tools contribute to data handling, transformation, or generation. This omission is consequential. AI-mediated processes can introduce errors that evade conventional editorial checks, including subtle misclassification, synthetic augmentation without clear labeling, or plausibility distortions that remain internally coherent while being externally implausible [18].

The absence of explicit governance in this domain also amplifies cross-journal variability. Where policies defer to umbrella guidance without local specification, expectations for provenance documentation remain implicit or inconsistently enforced. As AI-assisted data work becomes routine, governance frameworks that treat disclosure as sufficient risk mislocating the integrity problem: the central vulnerability often lies not in authorship transparency but in traceability.

Relatedly, journal policies rarely engage statistical plausibility as an explicit object of governance, despite the growing role of AI in analysis, interpretation, and synthesis. Where AI shapes analytic choices (whether through automated coding, model selection, result summarization, or narrative framing), traditional peer review practices may fail to surface errors, particularly when outputs appear polished and internally consistent [9].

This creates a structural vulnerability. Journals may enforce strong rules at the level of attribution while offering little guidance, standardization, or checkpointing for AI-assisted inference [1, 11]. In the absence of clearer expectations around analytic transparency, plausibility checks, and documentation of computational workflows, governance regimes risk being robust where verification is straightforward and thin where verification is more difficult, but epistemic stakes are higher [9, 11].

Peer review remains the least governed domain of AI use [1, 10]. Across many journals, reviewer-facing guidance is either absent or limited to confidentiality cautions, such as warnings against uploading manuscripts into external tools. Few policies specify whether AI assistance is acceptable in review, under what conditions it may be used, whether its use should be disclosed, or how editors should interpret AI-assisted review reports [10].

This omission is consequential because peer review is the primary institutional mechanism for evaluating methodological

rigor, plausibility, and integrity. When reviewers use AI to summarize manuscripts, assess novelty, draft critiques, or structure feedback, often without guidance or disclosure, current policies provide little assurance that such use strengthens rather than dilutes evaluative rigor [10]. Moreover, uneven access to secure or institutionally supported tools may introduce new capacity asymmetries among reviewers, potentially widening existing disparities in scientific gatekeeping [19, 20].

Such failures are not hypothetical: in 2026 the *New England Journal of Medicine* retracted a clinical image after the authors disclosed using an AI tool to manipulate the figure [21, 22], a lapse identified through post-publication scrutiny rather than during peer review [23].

A further and emerging blind spot concerns the integration of AI tools into editorial workflows themselves. As AI-enabled systems are incorporated into manuscript management platforms to support triage, screening, or quality checks, governance questions arise that existing policies rarely address: what forms of editorial AI assistance are acceptable, whether such use should be disclosed, and how algorithmic outputs should be weighed relative to human judgment.

This gap is structural because it sits outside the traditional scope of author-facing policies and is often invisible to researchers and reviewers. Without explicit governance, editorial AI risks becoming an unacknowledged layer of decision support, variable across journals and publishers, without transparent standards for accountability, auditability, or bias management.

Regional & institutional asymmetries. The structural gaps identified above do not affect all journals equally. Beyond variation in policy scope or specificity, the landscape analysis reveals pronounced regional and institutional asymmetries in the capacity to articulate, disseminate, and operationalize AI governance in scientific publishing. These asymmetries shape not only what is formally regulated, but also who can reasonably be expected to locate, interpret, comply with, or enforce existing policies in practice.

One of the clearest manifestations of this unevenness is policy discoverability. High-impact and digitally native journals were more likely to maintain centralized, clearly labeled AI guidance that was readily accessible through author and reviewer instruction pages. By contrast, many regional and society-run journals exhibited fragmented or opaque governance structures: AI-related guidance was dispersed across multiple documents, embedded within general author instructions, deferred to umbrella organizations, or absent altogether. Policies that cannot be readily located, interpreted, or verified by authors and reviewers fail functionally as governance mechanisms, regardless of stated intent. In practice, inaccessible or ambiguous guidance generates uncertainty at the point of submission and review, weakening accountability even where normative commitments to transparency or integrity have been articulated.

These asymmetries create conditions for policy arbitrage—a dynamic, analogous to regulatory arbitrage in other governance contexts [24], whereby AI-assisted research may flow toward venues with less explicit, less enforceable, or less visible governance. This dynamic does not require intentional evasion. In a publishing ecosystem characterized by heterogeneous disclosure standards, inconsistent reviewer guidance, and variable enforcement capacity, authors navigating multiple submission pathways may reasonably gravitate toward journals where compliance requirements are unclear or perceived as lower-friction. Where oversight is patchy, risks are displaced across journals rather than mitigated, with venues lacking explicit AI policies becoming inadvertent conduits for work that would face higher scrutiny elsewhere. Over time, this dynamic risks stratifying the scientific record along governance capacity rather than epistemic quality.

Underlying these patterns are differences in institutional resources and editorial capacity. Developing, maintaining, and enforcing AI governance requires sustained investments of time, expertise, legal review, and technical infrastructure. Large publishing groups and well-resourced journals are more likely to possess these capacities, enabling them to issue detailed guidance and update policies as tools and practices evolve. Many regional, society-run, or independently managed journals operate under far tighter constraints. Reliance on umbrella guidance may therefore represent a pragmatic response rather than a lack of awareness or commitment. However, without local specification, such reliance can leave critical gaps in implementation, particularly where regional research practices, data sources, languages, or ethical considerations diverge from those assumed in globally dominant policy frameworks.

At the same time, some governance approaches implicitly assume access to secure, enterprise-grade AI tools to mitigate confidentiality or data leakage risks. Where such access is uneven, policies may inadvertently exacerbate existing inequities among authors and reviewers, privileging contributors from well-resourced institutions while raising participation barriers for those operating in lower-resource settings [10].

These findings indicate that AI governance in scientific publishing is uneven not only in scope, but also in feasibility. Differences in policy discoverability, enforcement capacity, and material resources shape how governance is experienced in practice and create conditions under which integrity risks are redistributed across the publishing ecosystem rather than systematically addressed. These asymmetries expose the limits of policy-centric approaches that treat governance primarily as rule articulation. They reinforce the need for evaluative frameworks that attend to institutional capacity alongside policy content.

AI governance readiness levels. The structural gaps and institutional asymmetries identified above raise a further analytical question: how can the capacity of journals and

publishing organizations to govern AI-mediated risks be assessed in a systematic and comparable way? Existing policy inventories are effective at documenting whether guidance exists, but they do not capture whether governance mechanisms are operational, aligned with where risks actually arise, or feasible given institutional constraints.

To address this limitation, we introduce the AI governance readiness levels (AGRL) as a diagnostic framework for characterizing governance maturity in scientific publishing. AGRL shifts the analytic focus away from the mere presence of policies toward institutional capacity: the extent to which governance mechanisms are articulated, implemented, and integrated across the publication lifecycle in response to AI-enabled research and review practices.

AGRL is conceptually adapted from technology readiness level frameworks used in engineering and technology development, but it is tailored to the distinctive challenges posed by generative AI in scientific publishing. Rather than assuming a single optimal governance model, the framework recognizes that journals differ substantially in mission, scale, risk exposure, and resources. AGRL therefore functions as a descriptive and comparative tool, not a ranking system or prescriptive roadmap.

The framework comprises five readiness levels (0–4), each representing a qualitatively distinct governance capacity. The levels are cumulative: higher levels assume the capabilities of lower ones while extending oversight to additional domains of risk. Progression reflects not only policy articulation, but also investment in editorial workflows, reviewer guidance, verification infrastructure, and coordination mechanisms. In this sense, AGRL captures what institutions are currently positioned to do, rather than what they ought to do in principle.

At **AGRL-0**, no explicit AI-related governance is articulated. Author and reviewer instructions make no reference to AI tools, disclosure expectations, or acceptable use boundaries, and no mechanisms for enforcement or dispute resolution are specified. Governance relies on informal norms or implicit reference to external standards. While not necessarily indicative of negligence, this absence leaves institutions without a shared framework for managing AI-related risks.

AGRL-1 reflects the dominant point of convergence across much of the publishing ecosystem. Journals prohibit AI authorship, emphasize non-delegable human accountability, and endorse disclosure of AI use in principle [1, 5]. Governance at this level is largely declarative and concentrated on text production and attribution, with limited guidance on verification, enforcement, peer review practices, or analytic integrity. AGRL-1 establishes a normative baseline but leaves many higher-risk domains unaddressed.

At **AGRL-2**, governance extends beyond authorship to require more granular disclosure of AI use in data handling, analysis, and figure generation. Expectations for reviewer conduct begin to emerge, particularly around confidentiality and disclosure of AI assistance. However, verification remains largely manual, enforcement is inconsistent, and systematic evaluation of bias, fairness, or analytical plausibility is uncommon. Governance continues to depend heavily on author honesty and reviewer vigilance.

AGRL-3 marks a shift from transparency to evaluation. Journals operating at this level would require explicit assessment of bias, fairness, and applicability in AI-assisted research, alongside systematic attention to data provenance and analytical validity. Governance mechanisms extend to training reviewers and editors to identify AI-mediated risks and to adjudicating disputes related to algorithmic integrity. While technical and methodological tools to support this level exist [11], they are rarely embedded in routine journal workflows.

At **AGRL-4**, governance becomes systemic and reflexive. Oversight extends beyond individual journals to include coordinated post-publication monitoring, protected challenge and whistleblowing pathways, shared verification infrastructure, and cross-institutional learning. Equity considerations are integral to governance design, recognizing differential access to AI tools and resources [10]. Governance at this level evolves in response to empirical evidence and community input rather than static policy cycles.

To classify each journal, we mapped the six policy dimensions assessed in Table II against the AGRL level definitions above. Journals were assigned to AGRL-0 if no explicit AI policy was retrievable, AGRL-1 if they explicitly prohibited AI authorship and required general disclosure but lacked peer review guidance or enforcement. Journals were assigned to AGRL-2 only when all three elements were present simultaneously: granular disclosure, reviewer conduct guidance, and an explicit enforcement mechanism. Journals meeting only some of these criteria were classified as AGRL-1–2. No journal met AGRL-3 or AGRL-4 criteria. Table III presents the resulting classifications.

As shown in Table III, applying AGRL to the journals examined in this study reveals pronounced clustering at the lower end of the readiness spectrum. Most journals operate at AGRL-1, having established authorship prohibitions and disclosure norms without corresponding capacity to govern AI use in analysis, peer review, or post-publication oversight. A smaller subset—including JKMS and The Lancet—shows partial movement toward AGRL-2 through more detailed disclosure requirements, stricter scope restrictions, or explicit enforcement mechanisms, but does not fully satisfy all AGRL-2 criteria simultaneously. PLOS, JAMA and JMIR most closely approach AGRL-2, with granular disclosure requirements and structured reviewer guidance. No journal in the sample fully operationalizes AGRL-3 or AGRL-4.

This distribution should not be interpreted as institutional failure. Rather, it reflects the early and transitional nature of AI governance in scientific publishing, as well as the reality that higher readiness levels require resources, expertise, and coordination that exceed the capacity of most journals acting in isolation. In this sense, AGRL provides a diagnostic snapshot of current governance maturity rather than a linear

TABLE III: AGRL classification of sampled journals and umbrella organizations.

Journal/Organization	AGRL level	Key observables supporting classification
Clinics	0	No explicit journal-level AI policy located; authorship prohibition and disclosure only implied through ICMJE-aligned guidance; no AI-specific peer review guidance; no explicit enforcement.
BJMBR	0	No explicit journal-level AI policy located; authorship prohibition and disclosure only implied through ICMJE-aligned guidance; no AI-specific peer review guidance; no explicit enforcement.
NEJM	1	Explicit authorship prohibition; general disclosure required; writing addressed; peer review restricted without structured guidance; enforcement not explicit
SAMJ	1	Explicit authorship prohibition; general disclosure required; peer review AI use addressed, including disclosure expectations; no enforcement mechanism
CMJ	1	Explicit authorship prohibition; general disclosure required; no peer review guidance; no enforcement
PAMJ	1	Explicit authorship prohibition; general disclosure required; peer review prohibited; enforcement inherited from external standards
ICMJE	1	Explicit authorship prohibition; disclosure required; no peer review guidance; umbrella recommendations only
COPE	1	Explicit authorship prohibition; disclosure required; peer review confidentiality noted; umbrella position statement only
JKMS	1–2	Explicit authorship prohibition; disclosure required with specified format; no peer review guidance; explicit enforcement (rejection or retraction)
The Lancet	1–2	Explicit authorship prohibition; writing restricted; figures prohibited; peer review addressed; explicit enforcement; granular disclosure not specified
PLOS (Med/ONE)	2	Explicit authorship prohibition; granular disclosure (tool name, use description, validity evaluation); peer review prohibited; explicit enforcement
JMIR	2	Explicit authorship prohibition; most granular disclosure (complete prompts and transcripts required); reviewer AI permitted with structured restrictions; explicit enforcement
JAMA	2	Explicit authorship prohibition; most granular disclosure (tool name, version, manufacturer, dates of use); reviewer AI use prohibited or tightly restricted in peer review; explicit enforcement

Note: AGRL-1–2 denotes journals that satisfy all AGRL-1 criteria and partially meet AGRL-2 criteria but do not simultaneously operationalize granular disclosure, reviewer conduct guidance, and enforcement. No journal in the sample met AGRL-3 or AGRL-4 criteria. Classification is interpretive and based on publicly accessible policy documentation at the time of capture; it is not independently validated.

pathway that all journals are expected to follow. Because the framework was derived from the same sample to which it is applied, this classification should be read as interpretive rather than independently validated; we return to this constraint in the Limitations.

By making visible the distance between existing policy mechanisms and the demands imposed by AI-enabled research and review, AGRL clarifies why checklist-based or purely declarative approaches risk providing a false sense of security, and why addressing higher-order risks will require coordinated, capacity-aware approaches that extend beyond individual institutions. AGRL does not, however, specify what a transition between levels requires in practice: the resources, personnel, sequencing, or institutional arrangements that would enable a journal to move from AGRL-1 to AGRL-2 will vary considerably by context and cannot be determined from a narrative review alone. Identifying where governance currently stands is a necessary precondition for that work; prescribing the transition is a distinct undertaking that lies beyond the scope of this analysis.

Why static policies are insufficient. AI governance in scientific publishing remains largely organized around static policy documents, an approach poorly matched to the dynamics of AI-enabled research and review. These documents attempt to regulate a rapidly evolving socio-technical system (one in which social practices and technological tools are mutually shaping) through episodic revision and fixed categories, introducing structural limitations that extend beyond any single policy choice. This critique is not a dismissal of disclosure itself. In resource-constrained settings that lack the infrastructure for independent verification, disclosure remains a rational, low-cost, and evidence-based first step toward transparency, and may be the most feasible safeguard available; our concern is with treating disclosure as sufficient rather than as a foundation on which further verification should be built.

First, the pace of AI development routinely outstrips journal policy cycles. New models, interfaces, and applications emerge on the order of months, whereas policies are typically revised episodically and often reactively. As a result, guidance frequently lags behind practice, addressing tools or behaviors that have already shifted. Even relatively rapid policy convergence struggles to remain aligned with the expanding scope of AI use across research workflows.

Second, static policies presume interpretive clarity that no longer holds. Disclosure requirements and prohibitions rely on shared understandings of what constitutes AI use and where boundaries lie between writing assistance, analysis, and decision support. In practice, these distinctions are increasingly porous. This ambiguity complicates compliance

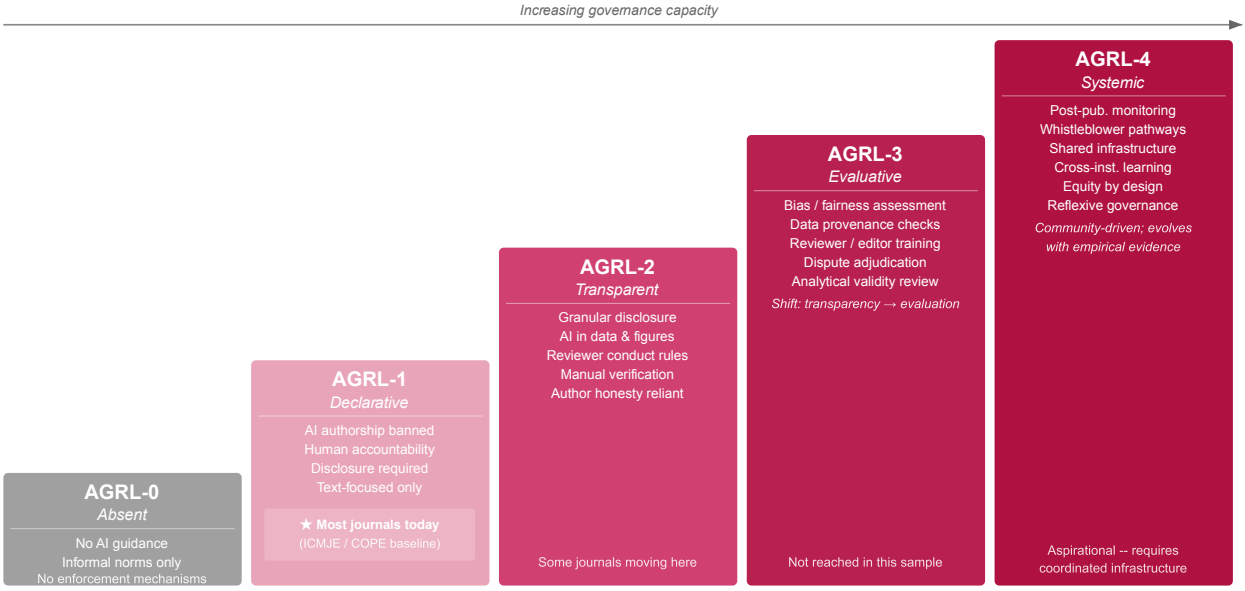

Figure 2: AI Governance Readiness Levels (AGRL): A diagnostic framework for assessing institutional capacity to govern AI across the scientific publishing pipeline. The framework comprises five cumulative levels (AGRL-0 through AGRL-4), each representing a qualitatively distinct institutional capacity to govern AI-mediated risks across the scientific publishing pipeline. AGRL-0 denotes the absence of any explicit AI governance; AGRL-1 (the dominant level in this sample) reflects declarative norms around authorship prohibition and disclosure; AGRL-2 extends governance to data handling, figures, and reviewer conduct; AGRL-3 shifts emphasis toward evaluative mechanisms including bias assessment, data provenance, and reviewer training; AGRL-4 represents systemic, reflexive stewardship with coordinated cross-institutional oversight and equity considerations integrated by design. Levels are cumulative: each higher level assumes all capabilities of lower levels. Applying these level definitions to the sampled journals, most journals cluster at AGRL-1, and no journal fully operationalises AGRL-3 or AGRL-4; because the framework was both derived from and applied to the same sample, this classification is interpretive rather than independently validated. Adapted from Technology Readiness Level frameworks [25]; AGRL is a descriptive and comparative diagnostic tool, not a prescriptive ranking.

and limits the ability of editors and reviewers to apply governance consistently.

Third, checklist-based approaches risk producing a false sense of security. Formal adherence to disclosure or authorship rules does not ensure that AI-assisted outputs are methodologically sound, unbiased, or appropriate to their research context. Where verification mechanisms, evaluative standards, and feedback loops are absent, governance may signal institutional values without meaningfully reducing epistemic risk.

These limitations are not failures of individual journals, but structural features of journal-centric governance. Scientific publishing operates through decentralized institutions with constrained editorial capacity, limited enforcement authority, and few mechanisms for continuous monitoring or coordinated adaptation. These constraints help explain why governance has gravitated toward declarative standards (such as authorship prohibitions) that are legally legible and straightforward to implement, while struggling to extend oversight into domains requiring ongoing evaluation, technical expertise, and institutional coordination. Addressing AI-mediated risks, therefore, demands a shift from rule articulation toward governance capacity building over time, from static compliance to reflexive stewardship.

Limitations. This analysis has several constraints. The journal sample, comprising 11 medical journals and two umbrella organizations, is small and purposively selected; a broader audit might yield different granular findings, particularly with respect to emerging or regionally specific governance practices. However, the structural patterns identified, including the concentration of governance at AGRL-1, persistent gaps in peer review and data integrity oversight, and pronounced regional asymmetries in policy discoverability, are sufficiently consistent across the sample to suggest they reflect ecosystem-level features rather than sampling artefacts. Policy capture depended on publicly accessible documentation at a fixed point in time, and governance language that is informally maintained or updated after the collection window may not have been captured. The AGRL framework was developed from and applied to the same dataset; independent validation in larger and more systematically sampled studies would strengthen its generalizability. These constraints reinforce the exploratory and diagnostic intent of the analysis. A further constraint concerns the regional tier specifically: the non-Western journals sampled are predominantly English-language, internationally indexed publications oriented toward global readership, and are therefore among the regional journals most likely to have already aligned with ICMJE and COPE norms. A sample including domestic-language journals serving primarily national clinical communities might reveal

different governance clustering; notably, no Spanish-language Latin American journal was included.

A further limitation concerns the PRAIDE framework itself, which is a theoretical, narrative proposal rather than an empirically validated protocol. We do not estimate the operational implications of its additional checkpoints, including the number of steps to completion, reviewer time and fatigue, personnel, or cost, as any such estimates would be speculative and could not be verified prior to implementation. We further acknowledge that, absent adequate support, additional review requirements could prove double-edged: reviewers facing a greater workload may feel more compelled to delegate evaluation to AI tools, the very behaviour the framework seeks to govern. Prospective evaluation of these trade-offs, alongside the distribution of checkpoints across multiple actors rather than the reviewer alone, remains an essential direction for future work.

From policing authors to stewarding science

This gap reflects not institutional failure, but a deeper mismatch between inherited models of editorial governance and the realities of algorithmically mediated science. Addressing this mismatch requires reframing the problem. AI governance in publishing cannot be reduced to policing individual authors or enforcing compliance with static rules. It instead calls for stewardship of the scientific knowledge system as a whole.

Journals remain indispensable actors in this effort, but they are insufficient as sole stewards. Historically, journal oversight has been designed to evaluate discrete manuscripts rather than dynamic, distributed processes spanning research design, data generation, analysis, review, and post-publication use. Editorial policies have therefore tended to address visible, well-defined risks, while more diffuse or systemic threats become legible only after failures surface. This reactive posture is not unique to AI. The replication crisis, large-scale data fabrication, and the rise of predatory publishing all emerged in environments where journals lacked both the mandate and the tools to monitor systemic risk in real time. AI intensifies these limitations by embedding automation across multiple stages of knowledge production, often beyond the direct visibility of editors and reviewers.

Journal-centric governance also reflects and reproduces existing asymmetries in capacity and influence. Dominant policy frameworks, typically developed by well-resourced publishers and umbrella organizations, often presume access to secure infrastructure, paid tools, and editorial bandwidth that does not generalize globally. Where such assumptions are embedded in governance design, they risk stratifying participation in scientific publishing along institutional and geographic lines, shifting burdens onto those least able to absorb them [10]. Moreover, journals operate as decentralized entities with limited mechanisms for coordination. They cannot pool resources for shared verification infrastructure, enforce standards beyond their own submission pipelines, or prevent policy arbitrage across venues. Even robust policies at individual journals therefore struggle to protect the integrity of the scientific record at the system level.

These constraints suggest that journals are best understood not as sovereign governors of AI-mediated science, but as essential nodes within a broader governance ecosystem. If journals cannot govern AI alone, governance must be understood as distributed and collectively enacted. Authors, reviewers, editors, institutions, funders, professional societies, and readers all participate, explicitly or implicitly, in shaping how AI is used, evaluated, and normalized within scientific practice. Governance is expressed not only through formal policies, but through everyday decisions that accumulate into norms. Within such a system, reflexivity becomes central: the capacity to observe consequences, learn from failures, and revise practices in response to evidence. Integrity failures, whether statistical anomalies detected post-publication or inequities revealed through downstream use, are not aberrations to be managed away, but signals that governance itself requires adaptation.

Collective stewardship also demands pluralism. Different disciplines, regions, and research contexts face distinct AI-related risks and operate under different constraints. Effective governance does not require uniform standards imposed from above, but shared principles that allow for local interpretation and adaptation. Stewardship in a low-resource clinical journal will look different from stewardship in a computational methods journal or a social science venue, yet each can be legitimate when aligned with a common commitment to integrity, accountability, and harm reduction.

Within this reframing, PRAIDE functions as an illustrative governance architecture rather than a prescriptive solution. By mapping oversight across the publication lifecycle, it makes visible where responsibility is currently concentrated and where it becomes diffuse, deferred, or displaced. Its value lies not in dictating what journals must do, but in enabling more precise questions about where governance exists, where it does not, and what forms of coordination are required. The framework highlights a central implication of stewardship: no single intervention is sufficient. Disclosure without verification, peer review without methodological support, or post-publication correction without monitoring infrastructure each address only fragments of a systemic problem. Integrity emerges from coordinated safeguards across stages and stakeholders—an outcome that exceeds the capacity of individual journals acting alone.

The challenge posed by AI in scientific publishing cannot be met through mandates alone. It instead issues an invitation: to recognize AI governance in scientific publishing as a collective responsibility and to engage in stewardship of the knowledge commons with intentionality and care. For researchers, stewardship means treating disclosure not as bureaucratic compliance but as epistemic responsibility. For reviewers, it entails extending evaluative rigor to computational provenance, statistical plausibility, and algorithmic bias, supported by appropriate tools and

institutional backing. For editors and journals, it requires moving beyond isolated policy refinement toward coordination, resource-sharing, and investment in shared infrastructure. For funders, institutions, and professional bodies, it demands treating governance capacity as a public good, worthy of sustained investment, global inclusion, and shared ownership.

The governance challenge posed by AI is ultimately a challenge of collective action. The narrow consensus achieved around authorship and disclosure represents an important first response, but it is insufficient for the scale of transformation underway. What follows will determine whether scientific publishing evolves toward reflexive stewardship or remains locked in reactive cycles that address yesterday's risks while tomorrow's accumulate. The choice is not between governance and innovation, or between integrity and efficiency. It is between static control and adaptive care, between policing isolated actors and stewarding the systems through which knowledge is produced, evaluated, and trusted. That work cannot be deferred, delegated, or automated away. It must be undertaken collectively, continuously, and with humility about what remains unknown.

Conclusion

The rapid integration of generative artificial intelligence into scientific research and publishing unsettles long-standing assumptions about how integrity, accountability, and trust are maintained. Journals have responded with notable speed, converging on prohibitions against AI authorship and establishing disclosure norms within a remarkably short period. These developments represent a genuine governance achievement. Yet they remain insufficient for governing AI-mediated workflows that now span the full lifecycle of scientific knowledge production.

Our review of 11 medical journals and two umbrella organizations found existing governance narrowly concentrated on authorship prohibition and disclosure, while peer review, data integrity, enforcement, and equity remained systematically under-governed. Regional journals were particularly underserved, often deferring to umbrella guidance without local specification, with consequences not only for compliance but for equitable participation in the global scientific record.

To move beyond inventories of what policies exist toward an assessment of what institutions can operationalize, we introduced the AI governance readiness levels (AGRL). Applying this framework revealed pronounced clustering at AGRL-1: most journals have established authorship prohibitions and disclosure norms, but lack the capacity to govern AI use in analysis, peer review, or post-publication oversight. No journal in the sample fully operationalized AGRL-3 or AGRL-4. This is not a finding of institutional failure, but a diagnostic of structural distance between where governance currently concentrates and where AI-related risks increasingly arise.

These findings together call for a fundamental reorientation. AI governance in scientific publishing cannot be treated as a compliance problem solved through static rules or disclosure checklists. It instead requires collective, reflexive stewardship of the scientific knowledge system, recognizing that responsibility is distributed across authors, reviewers, editors, institutions, funders, and professional bodies. Effective stewardship demands shared infrastructure, coordinated oversight, and governance frameworks that evolve in response to evidence rather than react to crises. We do not specify here which institution should anchor that coordination, how it should reach decisions, or what funding model should sustain it. These questions require stakeholder deliberation and empirical groundwork that lie beyond the scope of a narrative review. Our intent is to set the agenda for that coordination—to identify what must be built and why—rather than to prescribe its institutional form.

Trust in science is not automatic; it is continuously earned. In an era where AI can generate plausible but fabricated results, amplify latent biases, and blur the boundary between human and machine contribution, trust becomes a collective responsibility. It cannot be delegated to journals alone, automated through detection tools, or guaranteed by disclosure statements. The question is no longer whether AI will transform research, it already has, but whether the governance systems we build will be adequate to that transformation. Answering it requires moving beyond what journals should do toward what must be built together.

Acknowledgments

LAC is funded by the National Institute of Health through DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701, the National Science Foundation through ITEST 2148451, and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2024-00403047). TML is funded by the consortium's owner institutions – the University of Bergen, Western Norway University of Applied Sciences, the Institute of Marine Research, the Norwegian School of Economics, and SIVA SF – together with competitive grants from SR Bank, DnB, Agenda Vestlandet, and Nora.fo.

Use of AI and AI-assisted technologies

In keeping with the disclosure standard this paper advocates, we report our own use of generative AI. No AI system meets authorship criteria or is listed as an author, and the authors take full responsibility for all content. We used Claude Opus 4.6 (Anthropic) for initial parsing of policy text in the Methods (all extractions manually verified against primary sources), and to generate Figure 1 (PRAIDE), particularly to structure and typeset it and to help draft its annotations from content we supplied. The research underlying the framework belongs directly to the authors: we determined its design and content and, in selecting the illustrative case study, reviewed AI-suggested examples, rejected several, and independently identified and confirmed the case presented. AI-assisted tools may also have been used by authors for language editing of their own contributions; all such text was reviewed and revised

by the authors. All factual claims and citations were verified against primary sources.

Positionality

We note our position in relation to this work. PRAIDE and the AI governance readiness levels are frameworks we developed, and we have a stake in their reception. We have therefore framed both as descriptive and illustrative rather than validated or prescriptive, and we encourage independent scrutiny, adaptation, and empirical evaluation by others. Our authorship spans high-income and lower-resource research settings, which informs our emphasis on equity and feasibility, while also shaping the perspectives we bring.

Citation

Abulibdeh et al. From Policing Authors to Stewarding Science: Governing Artificial Intelligence Across the Scientific Publishing Pipeline in Medical Journals. *MIT Science Policy Review* **7**, 5-20 (2026). https://doi.org/10.38105/spr.cnmdfm0xb7.

Open access

References

[1] Ganjavi, C. *et al.* Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. *BMJ* **384**, e077192 (2024). 10.1136/bmj-2023-077192.

[2] Wang, Z. & Gong, M. A cross-disciplinary analysis of AI policies in academic peer review. *Learned Publishing* **39**, e2035 (2026). Published online 23 December 2025, 10.1002/leap.2035.

[3] Bagenal, J. Generative artificial intelligence and scientific publishing: urgent questions, difficult answers. *The Lancet* **403**, 1118–1120 (2024). 10.1016/S0140-6736(24)00416-1.

[4] Bhavsar, D. *et al.* Policies on artificial intelligence chatbots among academic publishers: a cross-sectional audit. *Research Integrity and Peer Review* **10**, 1 (2025). 10.1186/s41073-025-00158-y.

[5] International Committee of Medical Journal Editors (ICMJE). Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. PDF document (2026). Updated January 2026. Online: https://www.icmje.org/icmje-recommendations.pdf.

[6] Committee on Publication Ethics (COPE). Authorship and AI tools (2023). Online: https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools.

[7] Flanagin, A. *et al.* Reporting Use of AI in Research and Scholarly Publication—JAMA Network Guidance. *JAMA* **331**, 1096–1098 (2024). 10.1001/jama.2024.3471.

[8] Luo, X. *et al.* Reporting guideline for the use of Generative Artificial intelligence tools in MEdical Research: the GAMER Statement. *BMJ Evidence-Based Medicine* **30**, 390–400 (2025). 10.1136/bmjebm-2025-113825.

[9] Bedi, S. *et al.* Testing and evaluation of health care applications of large language models: A systematic review. *JAMA* **333**, 319–328 (2025). 10.1001/jama.2024.21700.

[10] Perlis, R. H. *et al.* Artificial Intelligence in Peer Review. *JAMA* **334**, 1520–1522 (2025). 10.1001/jama.2025.15827.

[11] Moons, K. G. *et al.* PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 (2025). 10.1136/bmj-2024-082505.

[12] Frontiers. Unlocking AI's Untapped Potential: Responsible Innovation in Research and Publishing. White paper, Frontiers Media. Survey of 1,645 researchers, May–June 2025 (2025). Online: https://www.frontiersin.org/documents/unlocking-ai-potential.pdf.

[13] AlFayyad, I., Zeegers, M. P., Bouter, L., Macdonald, H. & Schroter, S. Authors self-disclosed use of artificial intelligence in research submissions to 49 biomedical journals: A cross-sectional study. *medRxiv* (2025). Preprint, posted 25 October 2025, 10.1101/2025.10.24.25338574.

[14] Yoo, J.-H. Defining the boundaries of AI use in scientific writing: A comparative review of editorial policies. *Journal of Korean Medical Science* **40**, e187 (2025). 10.3346/jkms.2025.40.e187.

[15] Macdonald, H. & Abbasi, K. Riding the whirlwind: BMJ's policy on artificial intelligence in scientific publishing. *BMJ* p1923 (2023). 10.1136/bmj.p1923.

[16] Soulière, M. & Zhou, H. Emerging AI dilemmas in scholarly publishing. Topic Discussion, COPE Forum, Committee on Publication Ethics (2025). Forum held 1 July 2025. Online: https://publicationethics.org/topic-discussions/emerging-ai-dilemmas-scholarly-publishing.

[17] Drozdz, J. A. & Ladomery, M. R. The peer review process: Past, present, and future. *British Journal of Biomedical Science* **81**, 12054 (2024). 10.3389/bjbs.2024.12054.

[18] Gartlehner, G. *et al.* Artificial intelligence–assisted data extraction with a large language model: A study within reviews. *Annals of Internal Medicine* **178**, 1763–1771 (2025). 10.7326/ANNALS-25-00739.

[19] von Wedel, D. *et al.* Affiliation bias in peer review of abstracts by a large language model. *JAMA* **331**, 252 (2024). 10.1001/jama.2023.24641.

[20] Arno, A. *et al.* Accuracy and efficiency of machine learning–assisted risk-of-bias assessments in "Real-World" systematic reviews: A noninferiority randomized controlled trial. *Annals of Internal Medicine* **175**, 1001–1009 (2022). 10.7326/M22-0092.

[21] Wang, Y. & Mu, X. Bronchial Casts from Inhalation of Forest-Fire Smoke. *New England Journal of Medicine* **394**, 1634 (2026). Images in Clinical Medicine; published 18 April 2026; retracted, 10.1056/NEJMicm2518379.

[22] Wang, Y. & Mu, X. Retraction: Bronchial Casts from Inhalation of Forest-Fire Smoke. *New England Journal of Medicine* (2026). Retraction notice; published 29 April 2026, 10.1056/NEJMc2605962.

[23] Marcus, A. NEJM retracts case study for AI-manipulated imagery. Retraction Watch (2026). Accessed 29 May 2026. Online: https://retractionwatch.com/2026/05/01/nejm-retracts-case-study-for-ai-manipulated-imagery.

[24] Bartley, T. Transnational governance as the layering of rules: Intersections of public and private standards. *Theoretical*

Inquiries in Law **12**, 517 (2011). Online: https://doi.org/10.2202/1565-3404.1278.

[25] Mankins, J. C. Technology readiness levels: A white paper (1995).