# Fault-Based Publication Ethics:

## The Case for Source Custody

## in an Era of AI Citation Contamination

**Basil C. Puglisi, MPA**

A Human-AI Collaboration

Independent Practitioner-Researcher, basilpuglisi.com

**Working Paper | June 2026**

**Keywords:** fault-based publication ethics, source custody, citation contamination, hallucinated citations, fabricated citations, reference rot, due diligence evidentiary record, publication integrity, retraction, Source Provenance Ledger

## Abstract

Fabricated citations in the biomedical literature increased approximately tenfold from 2023 to early 2026, reaching one in 277 papers (Topaz et al., 2026), and 98.4% of flagged papers remained in the literature without correction at the time of the audit. Manual enforcement has failed. The response is automation: citation-verification pipelines, AI-detection products, and platform-level sanctions triggered by algorithmic flags are now entering scholarly production workflows. Automated enforcement solves the volume problem but creates a new one. AI-detection tools carry documented false-positive rates, and those false positives fall disproportionately on non-native English speakers, early-career researchers, and scholars at under-resourced institutions. An author wrongly flagged for citation fabrication under the current framework has no standardized evidentiary record to produce in rebuttal, because no publication ethics system asks authors to document what they verified at the time of use or provides a mechanism for producing that documentation when a citation is challenged. This paper argues that publication systems need a fault-based integrity model that makes verification effort visible, and it makes five contributions toward that model: (a) it defines four post-publication failure modes (retraction, link rot, content drift, and exposed fabrication) that can invalidate a source after good-faith use; (b) it documents the laundering mechanism through which invalidated sources compound in downstream work, drawing on evidence that over 94% of post-retraction citations in biomedicine never mention the retraction (Hsiao & Schneider, 2021; Bakker et al., 2024); (c) it establishes that reference rot predates AI and affects 13% to 75% of web citations depending on field and age (Klein et al., 2014; Zittrain et al., 2014), positioning AI as an accelerant on an already unstable citation substrate; (d) it proposes a five-level fault ladder separating fabrication, failure to verify, negligent verification, downstream contamination, and source decay, with graduated consequences calibrated to each level; and (e) it proposes a Source Provenance Ledger, a private evidentiary record maintained by the author and producible on challenge, building on Glynn's (2025) full-text reference deposit proposal and following established precedents for voluntary documentation in law, medicine, journalism, and financial services. Adoption is distributed across three channels: author prerogative (the author decides which outputs warrant documentation), publisher requirement (journals may require that submitters maintain a ledger producible on request), and insurance condition (errors-and-omissions policies may condition coverage on source custody records). The paper does not excuse careless AI use and does not claim the ledger is validated. It presents a fault-based framework, an evidence base, a staged validation roadmap, and an open invitation for replication. All referenced governance frameworks are published open-source under Creative Commons at github.com/basilpuglisi/HAIA.

## 1. Introduction: The Legibility Problem

Manual enforcement of citation integrity has failed. A peer-reviewed audit of 2.5 million biomedical papers found 4,046 fabricated references across 2,810 papers, with the rate increasing approximately tenfold from 2023 to early 2026 (Topaz et al., 2026). At the time of the audit, 98.4% of those flagged papers remained in the literature without correction (The Scientist, 2026). Peer reviewers do not catch fabricated citations: 76.7% report that they do not thoroughly check references, and 74.5% view peer review as ineffective at catching citation metadata errors (Xu et al., 2026, preprint). The human verification chain is broken at every level, and the field knows it.

The response is automation. Citation-verification pipelines such as CITADEL, CiteVerifier (from the GhostCite study), and commercial citation-checking products are entering scholarly workflows. arXiv's computer science section announced a one-year submission ban triggered by "incontrovertible evidence" of unchecked generative AI output, including hallucinated references and residual LLM artifacts (Nature, 2026; Research Information, 2026). No official arXiv policy page containing this ban language was located during verification; the enforcement standard appears to rest on section-level clarification reported through multiple independent news sources. Publishers are exploring automated reference-checking at the submission stage. The direction is clear: enforcement that humans failed to perform manually is being delegated to machines.

Automated enforcement solves the volume problem and creates a new one. AI-detection tools carry documented false-positive rates between 1% and 4% in controlled settings, with substantially lower accuracy in real-world conditions. Those false positives do not fall evenly. Non-native English speakers, neurodivergent writers, and researchers with naturally structured prose face higher rates of false accusation because their writing patterns overlap with the features detectors flag . Retraction impacts fall disproportionately on early-career researchers, particularly in developing countries. In one documented case, a researcher discovered an entire AI-generated article falsely published under his own name, a form of false attribution that citation-verification systems are not designed to catch (Jibril, 2025).

The prescribed remedy in the existing literature is already specific: falsely accused authors should collect drafts, revision history, and timestamps and use institutional due-process channels to appeal, supported by transparent policies that include human review and appeal processes (Undetectable.ai, 2025; ResearchGate, 2024). That remedy points toward source custody by another name. The field has diagnosed the need for contemporaneous documentation that an author can produce when challenged. It has not yet built the framework or the tooling.

This paper addresses that gap. The core problem is legibility. The current system makes verification effort invisible. An author who checked every citation against the primary record, ran Crossmark and Retraction Watch queries, and archived the source through the Internet Archive occupies the same position as an author who pasted unchecked LLM output, because no framework asks whether the author verified and no mechanism exists to record or produce the answer. A 2026 review of AI policies at the twelve largest

academic publishers found that nearly all lack explicit enforcement strategies, with "most onus and responsibilities on the researchers through self-reporting and trust-based mechanisms" (Sousa et al., 2026). When documented diligence and undocumented practice are treated the same, the incentive to verify collapses to personal conscience, which the behavioral evidence shows is insufficient (Xu et al., 2026, preprint: 41.5% of researchers copy-paste citations without checking). The Source Provenance Ledger proposed in this paper makes verification effort visible. Defense against false accusation, deterrence against carelessness, and fair adjudication of disputed citations all follow from legibility.

### 1.1 Contribution and Scope

This paper proposes a fault-based publication ethics model that preserves strong enforcement against fabrication and failure to verify while providing a framework for graduated response when source failures arise from circumstances beyond the author's reasonable control, and a private evidentiary record that authors can produce when their citations are challenged. The paper makes five contributions. It defines four post-publication failure modes that can invalidate a source after good-faith use (Section 2). It documents the laundering mechanism through which contaminated citations propagate in downstream scholarship (Section 3). It establishes the pre-AI baseline of reference rot and source decay, positioning AI as an accelerant (Section 4). It proposes a five-level fault ladder with graduated consequences (Section 7). It also proposes a Source Provenance Ledger, a private evidentiary record maintained by the author and producible on challenge, building on Glynn's (2025) full-text reference deposit proposal (Section 9).

Three research questions guide the investigation:

1. What distinguishes a source status change after use from author negligence, and can the distinction be operationalized in a publication ethics framework?
2. How do invalidated citations propagate through downstream scholarship, and at what rates does the status change travel with the citation?
3. What would a source-custody record need to contain for an author to produce evidence of diligence when a citation is challenged or flagged?

The paper does not claim the Source Provenance Ledger is validated or that the fault ladder has been adopted by any publisher. It presents a framework, an evidence base, a staged validation roadmap, and an invitation for the publication ethics community to test, challenge, or improve the proposal.

## 2. The Four Post-Publication Failure Modes

A source that supported a claim at the time an author used it can become invalid through four distinct pathways. The pathways differ in origin, detectability, and the degree to which the citing author could have prevented or anticipated the failure. These distinctions matter because an enforcement system that treats all four identically penalizes documented good faith at the same rate as willful fabrication.

**Retraction.** The cited paper existed, the author read it, the findings appeared on the page, and the citation accurately represented the content. After publication, the cited paper is

retracted for fabricated data, methodological failure, or editorial misconduct. The citing author had no access to the retraction investigation at the time of use and no mechanism for automatic notification after the retraction was issued. The cited source existed and was valid when accessed. It became invalid through an action taken by the original publisher.

**Link rot.** The cited URL resolved to a live page at the time of access. After publication, the domain expired, the hosting organization restructured its web presence, or the page was removed for editorial, legal, or administrative reasons. The URL now returns an error. The citing author cannot produce the content they originally referenced, and the reader cannot verify the claim. The source existed. The access path failed.

**Content drift.** The URL still resolves, but the content behind it has changed. The page no longer contains the passage the author cited, or the data have been updated, revised, or replaced without a formal correction notice. From the reader's perspective, the citation appears to point to a source that does not support the claim. From the author's perspective, the source supported the claim when accessed. The access path survived. The content did not.

**Exposed fabrication.** The citation may have been hallucinated by a large language model or introduced through another unverifiable generation process. The referenced paper does not exist, the author names are fabricated, the journal title may be real but the article is not, and the DOI, if provided, either resolves to a different paper or resolves to nothing. This failure mode differs from the previous three because the source never existed. It entered the citing paper because an AI tool generated a plausible-looking reference that the author, or a prior author whose work was cited, did not verify against the primary record. Ansari (2026) classifies the subtypes of this failure mode across a five-category taxonomy: total fabrication, partial attribute corruption, identifier hijacking, placeholder hallucination, and semantic hallucination. The taxonomy matters because partial attribute corruption and identifier hijacking produce citations that superficially pass verification checks, making them harder for a diligent author to catch than total fabrications.

The enforcement question at the center of this paper turns on which failure mode is present. An author whose citation failed through retraction or link rot stood in a different evidentiary position at the time of use than an author whose citation was fabricated by an unchecked AI tool. Current publication ethics does not make that distinction.

## 3. The Laundering Mechanism

The concern about citation contamination is that false or invalidated citations do not stay where they land. They propagate through downstream work because later authors treat the citing document as if it has already performed verification and stop checking behind it. The evidence for this mechanism is strongest in the retraction context, where the pattern has been measured at scale, and is emerging in the AI-fabrication context, where the first empirical observations of citation cascading have appeared.

In biomedicine, the most comprehensive evidence comes from a database-wide analysis of 7,813 retracted papers indexed in PubMed, covering 169,434 citations and 48,134 citation contexts across six decades. Retracted papers continued to be cited after retraction, and the retraction did not change how subsequent authors used them. Among 13,252 post-retraction citation contexts, only 722 (5.4%) acknowledged the retraction (Hsiao & Schneider, 2021). The remaining 94.6% of citing authors treated the retracted paper as valid scholarship. A separate analysis of retracted publications in evidence synthesis found that over 94% of post-retraction citations in biomedicine did not mention the retraction, and that retracted data were entering systematic reviews, the evidence tier that clinicians and policymakers rely on for treatment guidelines and regulatory decisions (Bakker et al., 2024). Avenell et al. (2019) found that 41% of systematic reviews citing a set of retracted clinical trials would likely change their findings if the retracted data were removed.

The COVID-19 pandemic produced a real-time case study of this mechanism under conditions of maximum urgency and minimum verification time. An analysis of 212 retracted COVID-19 articles found that they received 1,036 citations, 80% of which occurred after retraction, and 86% of the citing papers did not indicate the retraction status (De Oliveira Andrade, 2022). Clinicians making treatment decisions during the pandemic may have relied on evidence chains that included retracted findings without any signal that the underlying data had been withdrawn.

Retraction status often does not travel with the citation. The PDF of the retracted paper persists on personal computers, institutional repositories, and preprint servers that do not update their copies when the publisher issues a retraction notice. Authors who access the paper through these secondary channels have no reason to suspect the source is compromised.

In the AI-fabrication context, the first large-scale empirical evidence of citation cascading appeared in 2026. The GhostCite study identified instances in which the same fabricated citation appeared across up to 16 separate papers published at AAAI, IJCAI, and NeurIPS (Xu et al., 2026). The mechanism is straightforward: one paper contains a hallucinated reference generated by an LLM, and later authors copy the reference from that paper's bibliography without verifying it against the primary record. The fabricated citation acquires apparent legitimacy through repetition.

A position paper accepted at NeurIPS 2025 named this mechanism, in the authors' phrase, a "survey paper DDoS attack" on the research community, arguing that AI-generated survey papers "surface unverified or even hallucinated citations" that flood preprint platforms and "drown out genuinely insightful contributions" (Lin et al., 2025). While the paper did not quantify propagation rates, it identified the structural vulnerability: when reference lists are generated by LLMs and published without human verification, the contamination enters a system that has no automatic mechanism to quarantine it.

AI-generated citation fabrication enters a scholarly record that paper mills have already contaminated through systematic production of fake or low-integrity publications sold for a fee, a practice estimated to affect approximately 2% of the scientific literature in that

study (Parker et al., 2024). AI lowers the cost of that contamination pattern by generating plausible references, text, and evidence trails at scale.

The network-level dynamics of citation propagation have also been studied in the context of a single widely cited paper retracted from Nature. Van der Vet and Nijveen (2016) constructed the entire citation network and found that directly citing articles repeated the retracted result, while indirect citations (citations of citations) did not propagate the retracted finding in that case. This suggests that the primary contamination risk sits at the first layer of downstream citation, where authors copy from the original reference list, and that the risk may attenuate at greater distances from the contaminated source. The finding comes from a single case study and the authors acknowledge this limitation, but it aligns with the GhostCite observation that fabricated citations propagate through direct bibliography copying.

The practical implication is that the highest-risk contamination window may be narrow but consequential. If the status of a source could be checked and recorded at the moment of citation, the most damaging propagation channel, direct bibliography copying without verification, could be interrupted. That is the operational premise of the Source Provenance Ledger proposed in Section 9.

**Figure 1. The Contamination Lifecycle**

How a valid source becomes invalid after citation and compounds through downstream work

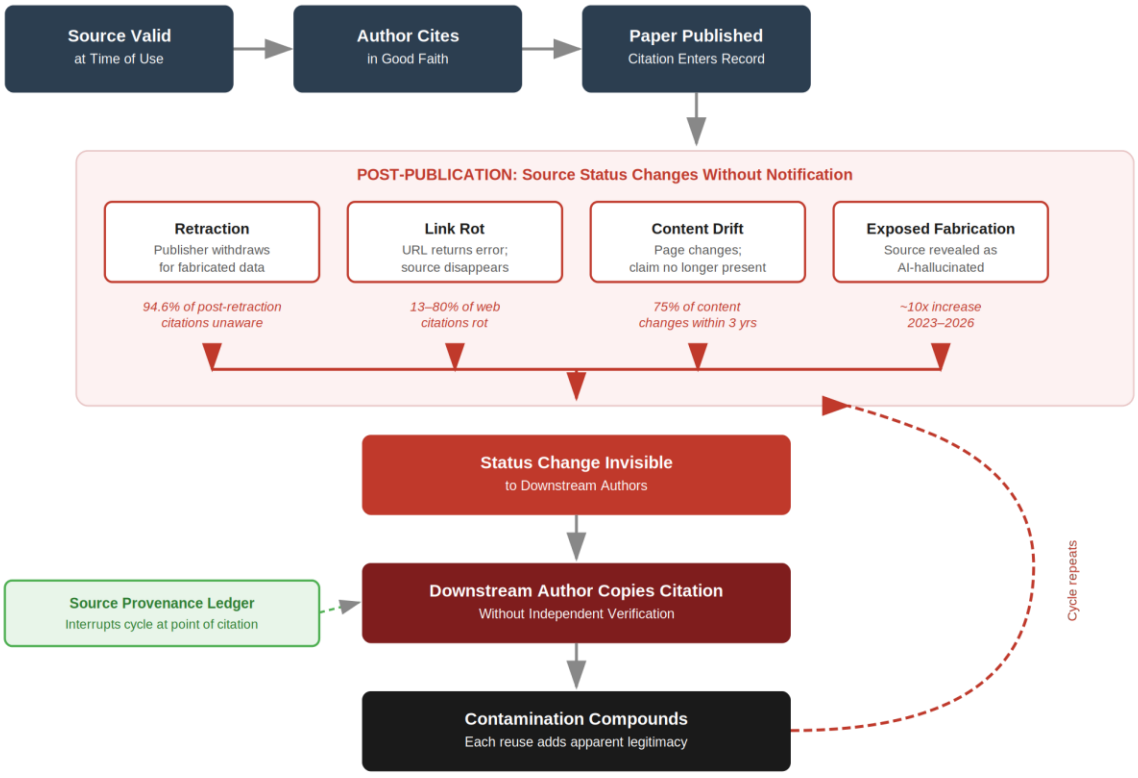



*Figure 1. The Contamination Lifecycle*

## 4. The Pre-AI Baseline: Reference Rot and Source Decay

The instability of cited sources is older than generative AI. Establishing this baseline matters for the paper's argument because it helps prevent the fault ladder from being dismissed as a response to an AI-specific problem. Source decay is a structural condition of web-based scholarship, and AI has accelerated it without creating it.

The foundational measurement comes from Klein et al. (2014), who examined over one million references across nearly 400,000 articles in three corpora: arXiv, Elsevier, and PubMed Central. For articles published in 2012, 13% of arXiv references, 22% of Elsevier references, and 14% of PubMed Central references suffered from link rot, meaning the URL no longer resolved to any content. For older articles published in 1997, the rates climbed to 34% for arXiv, 66% for Elsevier, and 80% for PubMed Central. The trend was consistent across all three corpora: reference rot increases with age, and recent articles are already affected at rates that compromise verification within a few years of publication.

The problem extends beyond the sciences. Zittrain, Albert, and Lessig (2014) investigated reference rot in legal scholarship across three Harvard law journals and found that between 65% and 73% of URI references suffered from link rot or content drift. In United States Supreme Court opinions, more than 50% of URLs no longer produced the information originally cited. The authors concluded that legal citation, which depends on the ability to verify the evidence underlying judicial reasoning, was being built on a medium that could not guarantee the persistence of that evidence. Their response was Perma.cc, a web archiving service that creates permanent snapshots at the time of citation, now adopted by over 120 law libraries and recommended by the Bluebook citation manual.

Content drift compounds the problem in ways that link rot does not. When a URL fails, the absence of content is visible: the reader encounters an error and knows the source is missing. When a URL resolves but the content has changed, the reader has no signal that the page no longer contains what the citing author referenced. Jones et al. (2016) found that as much as 75% of referenced content had changed to some degree within three years, and the overall combined rate of link rot and content drift reached 65% to 73% across the studied corpora.

An access date, the standard citation element intended to document when a web source was consulted, does not, by itself, solve the problem. The access date records when the author visited the page, but it does not preserve what the page said. It does not prove that the content supported the claim. And it provides no evidence of diligence beyond the fact that the author typed a date into a citation field.

The implication for the fault ladder is direct. Source decay exists as a baseline condition of digital scholarship. An author whose cited source later disappeared, changed, or became inaccessible is experiencing a failure mode that affects between one fifth and three quarters of all web citations within a few years, depending on the field and the age of the reference. Treating this author identically to one who fabricated a citation or pasted unchecked AI output conflates structural vulnerability with individual misconduct.

## 5. The Scale and Peer-Review Evidence

### 5.1 The Acceleration of Fabricated Citations

Two empirical studies, one peer-reviewed and one preprint with conference acceptance, provide the quantitative foundation for the paper's claims about the scale of AI citation contamination.

The Lancet study (Topaz et al., 2026) scanned 2,471,758 papers in PubMed Central's Open Access subset from January 2023 through February 2026, verifying 97.1 million references that carried PubMed identifiers against bibliographic records in PubMed and Crossref. The study identified 4,046 fabricated references across 2,810 papers. The fabrication rate held stable at approximately four citations per 10,000 papers throughout 2023, then rose sharply beginning in mid-2024, reaching approximately 57 per 10,000 papers by early 2026. The raw ratio of paper-level rates (one in 2,828 in 2023 to one in 277 in early 2026) represents an approximately tenfold increase; Retraction Watch characterized this as a "12-fold" increase based on a different denominator, which the paper adopts for consistency with the secondary literature. The authors attributed the surge to the proliferation of AI writing tools, noting that previous studies estimated 30% to 69% of LLM-generated references in biomedical contexts are fabricated. At the time of the audit, 98.4% of the flagged papers remained in the literature without correction (The Scientist, 2026). Enforcement had not reached the contaminated documents.

The GhostCite study (Xu et al., 2026, preprint, poster accepted at IEEE S&P 2026) analyzed 2.2 million citations from 56,381 papers published at top-tier AI, machine learning, and security venues from 2020 through 2025. The study confirmed that 604 papers (1.07%) contained invalid or fabricated citations, with an 80.9% year-over-year increase in 2025. A complementary LLM benchmark tested 13 models across 40 research domains and found hallucination rates ranging from 14.23% to 94.93%, with significant domain sensitivity.

A third quantitative signal, more conservative in absolute terms, comes from the SPY Lab at ETH Zurich (Tramèr, 2025, blog with open-source code). An analysis of arXiv references estimated that approximately 0.025% of references (roughly one in 4,000) appeared hallucinated, with a clear upward trend accelerating from early 2025. The author noted explicitly that this figure is "almost certainly a significant underestimate" because the detection pipeline only checked arXiv-to-arXiv title matches and missed non-arXiv fabrications, non-existent papers, and cases where the title was wrong but the DOI resolved.

### 5.2 Field-Specific Variation

The three estimates differ by more than an order of magnitude: 0.025% of references at SPY Lab's conservative arXiv floor, 1.07% of papers at AI/ML venues in GhostCite, and one in 277 papers (approximately 0.36%) in the Lancet's biomedical audit. These numbers are not contradictory because they measure different populations, use different detection methods, and define "fabricated" differently. The variation itself is a finding. A uniform enforcement standard applied across fields with contamination rates that differ by a factor of 14 to 40 may produce different collateral effects in different communities, and the

communities with higher baseline rates will be disproportionately affected by bright-line sanctions.

### 5.3 The Verification Chain Is Broken

The most consequential finding from the GhostCite study is behavioral. Through a survey of 94 valid responses, Xu et al. (2026, preprint) identified what the authors call a "verification gap" that operates at every level of the scholarly production chain. Among researchers, 41.5% reported that they copy-paste BibTeX entries without checking them against the primary source, and 44.4% reported taking no action when encountering a suspicious reference. Among peer reviewers, 76.7% do not thoroughly check references, 80% have never suspected a fake citation, and 74.5% view peer review as ineffective at catching citation metadata errors. The NeurIPS 2025 case confirms this at the highest selectivity tier: 100 fabricated citations across 53 accepted papers survived review by three to five expert researchers per submission (Ansari, 2026, preprint).

These numbers reframe the problem. The challenge is that AI tools produce fabricated citations, but the reason those fabrications survive is that the human verification chain does not check for them at any stage. Authors copy without verifying. Reviewers do not check references. Editors rely on a peer review process that 74.5% of its own participants view as ineffective at catching this specific failure. The Source Provenance Ledger proposed in Section 9 addresses this gap by moving verification from the assumed category to the documented category. When verification cannot be assumed at any gate, it must be recorded at the source.

### 5.4 Evidence Status

The argument in this paper does not depend on any single preprint. The following list classifies the major sources by evidence tier to ensure the reader can assess the weight each carries.

**Peer-reviewed:** Topaz et al. (Lancet, 2026); Hsiao & Schneider (QSS, 2021); Klein et al. (PLOS ONE, 2014); Zittrain et al. (Harvard Law Review Forum, 2014); Jones et al. (PLOS ONE, 2016); Avenell et al. (BMJ Open, 2019); Bakker et al. (BMJ Evidence-Based Medicine, 2024); Van der Vet & Nijveen (Research Integrity and Peer Review, 2016); Glynn (European Science Editing, 2025).

**Preprint with conference acceptance:** Xu et al./GhostCite (arXiv, poster at IEEE S&P 2026).

**Preprint:** Ansari (arXiv, 2026); Lin et al. (arXiv, NeurIPS 2025 poster).

**Blog with open-source code:** Tramèr/SPY Lab (ETH Zurich, 2025).

**Journalism:** Nature (2026); Research Information (2026); The Scientist (2026); Inside Higher Ed (2026); FAPESP (2022).

**Official policy:** COPE Position Statement (2023, updated 2025); COPE Glossary (2026).

The peer-reviewed reference rot and retraction-persistence studies establish the underlying structural problem. The preprints supply early measurement of an accelerating AI-specific failure mode. The journalism documents enforcement actions and expert commentary. The official policy documents are the standards the paper engages with.

## 6. Practitioner-Researcher Positionality

The frameworks and proposals in this paper emerged from sustained multi-platform AI collaboration practice, during which the citation verification problem documented in Sections 2 through 5 was encountered directly. Working across multiple AI platforms in structured parallel research, the practitioner documented multiple instances in which AI platforms generated citations that did not exist, attributed findings to wrong authors, or provided DOIs that resolved to different papers than the ones named. In one documented case, a four-of-six AI platform majority introduced citation errors that the human arbiter overrode, with subsequent review confirming the minority position was correct. The checkpoint-verification discipline that caught these errors, developed through operational experience in law enforcement, is the same discipline the Source Provenance Ledger proposes to formalize: every consequential action passes through a named human with binding authority at a defined checkpoint, verified against standards before proceeding.

This work is not affiliated with any corporation, funded by any investor, or peer-reviewed by a formal journal review process. It constitutes working concepts with documented practitioner evidence, not validated standards. The practitioner observations do not replace the empirical studies cited in Sections 3 through 5. They provide complementary first-person documentation of the problem this paper addresses. The governance frameworks referenced in this work, along with the author's full methodological background, publication history, and the multi-platform research methodology used to assemble this paper's evidence base, are described in Appendix C and published open-source under Creative Commons at github.com/basilpuglisi/HAIA.

## 7. The Fault Ladder

Current enforcement applies a uniform consequence to citation failures regardless of the author's conduct at the time of use. The evidence in Sections 2 through 5 shows that source failures exist on a spectrum of culpability, from deliberate fabrication to passive source decay, and that a uniform standard applied across that spectrum makes verification effort invisible, which removes any institutional incentive to verify.

### 7.1 The Gap Between COPE and arXiv

The Committee on Publication Ethics and arXiv occupy different positions on enforcement, and the gap between them is where the fault ladder fits.

COPE's Position Statement on Authorship and AI Tools states that authors are "fully responsible for the content of their manuscript, even those parts produced by an AI tool, and are thus liable for any breach of publication ethics" (COPE, 2023). COPE's Glossary,

updated April 2026, defines “author sanctions” specifically as “actions that editors or publishers might take in response to suspected or proven misconduct, such as banning authors from publishing,” and advises that “such author sanctions should not be imposed as they risk restricting academic freedom,” recommending instead “an educational approach” with “escalation to the relevant institution or regulatory authority if necessary” (COPE, 2026).

COPE’s posture is already graduated. The Position Statement establishes responsibility as a principle. The Glossary advises against the most severe penalty (banning) while preserving institutional escalation for genuine misconduct. COPE’s Core Practices and retraction guidelines add further granularity: corrections for the record, retractions for unreliable findings, expressions of concern for ongoing investigations, and educational engagement for authors whose errors do not constitute misconduct. Read together, COPE’s guidance assigns responsibility without prescribing a single undifferentiated sanction.

arXiv’s enforcement action operates differently. The reported one-year submission ban applies to all listed authors when a submission contains “incontrovertible evidence” of unchecked LLM output, with responsibility assigned “irrespective of how the contents were generated” (Nature, 2026; Research Information, 2026). The ban is the kind of blanket author sanction COPE’s Glossary advises against. It does not graduate by fault level. It does not distinguish fabrication from failure to verify from downstream contamination from source decay. After the ban period, subsequent submissions must first be accepted at a peer-reviewed venue.

The tension in the field is between COPE’s graduated philosophy and arXiv’s blunt enforcement instrument. COPE advises against bans as a general sanction. arXiv’s reported enforcement action imposes one. The fault ladder proposed here bridges this gap. It preserves arXiv’s enforcement at the levels where the author’s conduct supports stronger enforcement (fabrication and failure to verify) while honoring COPE’s anti-ban guidance at the levels where it applies (downstream contamination and source decay). An author who can produce evidence of verification occupies a different position than one who cannot, and the ladder makes that distinction visible to both the platform and the ethics body.

What remains unresolved in either framework is the evidentiary question, and the gap propagates industry-wide. A review of AI policies at the twelve largest academic publishers, which together account for two-thirds of all scholarly publications worldwide, found that most adopted COPE’s guidance and inherited its enforcement void, with policies focused on immediate procedural compliance and almost no consideration of enforcement mechanics or long-term accountability (Sousa et al., 2026). What standard of verification satisfies the responsibility COPE assigns? COPE does not specify what constitutes sufficient checking. arXiv does not ask whether the author checked. An author who verified every citation and archived the evidence is treated identically to one who never opened a source, because no mechanism exists to ask, record, or produce the answer. The Source Provenance Ledger fills that evidentiary gap.

### 7.2 Level 1: Fabrication

The author invents a source, knowingly submits a false reference, or creates citation metadata that does not correspond to any existing publication. This is research misconduct under every applicable framework, including COPE guidance, arXiv moderation standards, and the U.S. federal definition of fabrication under 42 CFR 93. The appropriate response is retraction, notification to the author's institution, and suspension of submission privileges. Evidence of diligence is irrelevant because the act is intentional.

### 7.3 Level 2: Failure to Verify

The author includes citations in a manuscript without verifying their existence against the primary record, and the unverified citations prove to be fabricated or nonexistent. The author did not intend to deceive but failed to exercise basic verification. This level applies regardless of whether the citation originated from an LLM, from another paper's bibliography, or from a colleague's recommendation. The distinguishing conduct is the absence of any verification, not the tool that produced the error.

The distinguishing evidence is the absence of any ledger entry. An author who cannot produce evidence of having accessed the cited source at all (no archived copy, no recorded passage, no access date beyond the citation field, no status check) has not verified. The conduct, not the tool, determines the level.

The legal parallel is instructive but not directly transferable. In *Mata v. Avianca, Inc.* (2023), the United States District Court for the Southern District of New York sanctioned attorneys who used ChatGPT-generated case citations without verification, finding that they had "abandoned their responsibilities." The court applied Federal Rule of Civil Procedure 11(b), which requires that an attorney certify, after "an inquiry reasonable under the circumstances," that factual contentions have evidentiary support. The legal analogy is offered as evidence that mature professional systems can distinguish responsibility for a filing from fault in the verification process.

It is not offered as a transplantable rule. Academic publishing lacks the adversarial procedure, contempt authority, and licensure mechanisms that give Rule 11 its enforcement power. The analogy illuminates the principle; the academic system must build its own mechanism.

The appropriate response at this level is correction, disclosure, and graduated sanction. A first offense with a small number of unverified citations warrants correction and warning. Repeated or extensive failures warrant stronger sanctions.

### 7.4 Level 3: Negligent Verification

The cited source exists but does not support the claim the author attributes to it. The author accessed the source but either misread it, misattributed its findings, or cited it without reading beyond the abstract or title. This is a verification failure, and the appropriate response is correction and disclosure. Sanctions are warranted only if the negligence is systematic.

The distinguishing evidence is a ledger entry that shows the author accessed the source and recorded a passage, but the passage does not support the claim as attributed. The entry documents engagement. The error is in interpretation, not in the absence of checking. Level 2 means the author never opened the source. Level 3 means the author opened it and got it wrong.

### 7.5 Level 4: Downstream Contamination

The author cited a real source that later proved contaminated, retracted, or partly fabricated. At the time of citation, the source had not been flagged by any correction notice, Retraction Watch, or publisher update. The author acted in good faith. The fault lies upstream, with the original fabricator, the publisher that failed to catch the problem, or the retraction system that failed to propagate the status change.

The appropriate response is disclosure and correction once the contamination is discovered. If the author can produce evidence that a status check was performed at the time of use and the source was clean, the author has documented good faith. The tools that make this recordable exist: Crossmark, Retraction Watch, Crossref metadata, and publisher correction notices.

### 7.6 Level 5: Source Decay

The source existed and supported the claim when the author used it. It later disappeared (link rot), changed (content drift), or became inaccessible through no action by the citing author. The source's absence after publication is a failure of the digital medium.

The appropriate response is correction or an updated link, accompanied by an editorial note explaining the change. No sanction and no finding of fault are warranted. An author who preserved the source at the time of use can produce contemporaneous evidence that the source existed and supported the claim.

### 7.7 Classification: A Decision Tree for Editors

The five levels are distinguished by the author's conduct. Three questions classify a citation failure:

**Question 1: Did the cited source exist at time of use?** If the source never existed (the citation is fabricated or hallucinated), classify as Level 1 if available evidence indicates the author knew, or Level 2 if the author failed to verify its existence. If the source existed, proceed to Question 2.

**Question 2: Did the author access and read the source?** If the author cannot produce any evidence of having accessed the source (no ledger entry, no archived copy, no recorded passage), classify as Level 2 (failure to verify). If the author accessed the source, proceed to Question 3.

**Question 3: Did the source support the claim attributed to it at the time of use?** If the source existed and the author accessed it but the source does not support the claim, classify as Level 3 (negligent verification). If the source supported the claim at time of use but later

became invalid, classify as Level 4 (downstream contamination, if the source was retracted or exposed as fabricated) or Level 5 (source decay, if the source disappeared or changed without a formal retraction).

**Worked examples:**

*Example 1:* A researcher includes a citation that resolves to a real journal but the specific article does not exist. The researcher cannot produce any evidence of having opened the article page, read a passage, or confirmed the content. Classification: Level 2 (failure to verify). Whether the citation originated from an LLM, from another paper's bibliography, or from a database error is immaterial. The conduct is the same: no verification performed.

*Example 2:* A researcher cites a 2024 clinical trial that was retracted in early 2026 for data fabrication. The researcher accessed the paper in mid-2025, read the findings, and cited them accurately. A ledger entry records the access date, the specific claim extracted, and a Retraction Watch check that returned no flag. Classification: Level 4 (downstream contamination). The author verified. The source status changed after verification.

*Example 3:* A researcher cites a government report hosted on an agency website. The report supported the claim at the time of citation. Eight months later, the agency restructures its web presence and the URL returns a 404 error. Classification: Level 5 (source decay). If the researcher archived the page via the Internet Archive, they can produce evidence that the source existed and supported the claim. If not, the absence of an archive is a missed opportunity for self-protection, not misconduct.

**Figure 2. The Fault Ladder**

Graduated consequences calibrated to the author's conduct at the time of citation

| | Level | Author Conduct | Decision Test | Appropriate Response | Ledger Evidence |
|---|---|---|---|---|---|
| SANCTION | 1 | **Fabrication** Author invents source or knowingly submits false reference | Source never existed. Author knew or created the fabrication. | Retraction. Institutional notification. Submission ban. | Irrelevant. Intentional act. Absence of any ledger entry is consistent with fabrication. |
| SANCTION | 2 | **Failure to Verify** Author includes citation without checking it exists. Tool is irrelevant. | Source did not exist, or defect was discoverable. Author never opened or read source. | Correction + disclosure. Warning (first offense). Graduated sanction (repeated failures). | No ledger entry exists. Author cannot produce evidence of having accessed the source. |
| CORRECTION | 3 | **Negligent Verification** Author accessed source but misread or misattributed findings. | Source exists but does not support the claim. Author accessed it (has ledger entry). | Correction + disclosure. Sanctions only if systematic negligence. | Entry exists with recorded passage. Interpretation was wrong, but engagement was real. |
| GOOD FAITH | 4 | **Downstream Contamination** Source later retracted or exposed as fabricated. | Source supported claim at time of use. Status changed after citation. Defect was not discoverable at time of use. | Disclosure + correction. No sanction if author can produce status check record. | Entry with Crossmark or Retraction Watch check showing source was clean at access date. |
| GOOD FAITH | 5 | **Source Decay** Source disappeared or changed after publication. Medium failed. | Source existed and supported claim. URL now returns error or content has changed. | Correction or updated link. Editorial note. No sanction. No fault. | Entry with archive link (Internet Archive or local copy) proving source existed and supported claim. |

*Figure 2. The Fault Ladder*

## 8. Governance Integration

The fault ladder addresses a normative gap: how should the enforcement response vary with the author's conduct? The Source Provenance Ledger proposed in Section 9 addresses an operational gap: how can an author document and preserve evidence of verification at time of use, and produce it when challenged? Before specifying the ledger, this section maps the operational gap against existing governance infrastructure.

### 8.1 Established Precedents for Private Production-on-Challenge

The concept of a private evidentiary record maintained by the practitioner and producible when challenged is not new. It follows a pattern established across multiple professions.

Attorneys maintain work product that is not submitted with every filing but can be produced to show the "inquiry reasonable under the circumstances" that Rule 11(b) requires. Physicians maintain clinical records motivated by professional duty and malpractice insurance: a practitioner who cannot produce records when challenged bears a heavier evidentiary burden. Journalists maintain source files protected by shield laws, produced when reporting is challenged, with the journalist deciding what to share subject to legal constraints. Financial services professionals maintain audit trails motivated by regulatory requirements and errors-and-omissions insurance: inability to produce records during an audit creates adverse inference.

In each case, the record is private, maintained as part of professional practice, and produced on challenge rather than submitted by default. The record's existence creates an evidentiary asymmetry between the practitioner who documented and the one who did not. No profession requires practitioners to submit their verification records with every deliverable. Every profession recognizes that the ability to produce records when challenged affects the outcome of disputes.

Publication ethics has no equivalent practice. Authors are told they are responsible for their citations, but no framework asks them to maintain documentation of what they verified, and no mechanism exists for them to produce that documentation when a citation is challenged or flagged by an automated system.

### 8.2 Existing Infrastructure the Ledger Integrates With

The ledger does not exist in a vacuum. Components of the verification and preservation infrastructure already exist, and the ledger's contribution is integration, not invention.

Crossmark, operated by Crossref, provides a standardized mechanism for readers to check the current status of a published article, including corrections, retractions, expressions of concern, and updates. It checks publisher-participating records but does not cover non-journal sources, preprints, or web content, and it does not record whether an author checked before citing.

The Retraction Watch database, available through Crossref, contains over 40,000 retractions searchable by author, journal, and reason. It is updated every working day. An

author can check whether a cited source has been retracted, but no submission system currently requires or records that the check was performed.

The Internet Archive's Save Page Now creates timestamped snapshots of web pages at the moment of access. It is free, universally accessible, and requires no institutional affiliation. Perma.cc, developed by the Harvard Library Innovation Lab, provides a more structured archiving service and has been reported as adopted by more than 120 law libraries and seven state court systems. The Bluebook citation manual recommends it as a "reliable web archiving service" (Zittrain et al., 2014; Dulin et al., 2017). Perma.cc access is institution-gated through subscribing libraries; researchers without an institutional affiliation should use the Internet Archive as the universally accessible alternative. The minimum viable ledger entry should not depend on tools that require institutional membership.

Glynn (2025) proposed the closest existing parallel to the Source Provenance Ledger: full-text reference deposit, requiring authors to submit the complete text of each cited source along with their manuscripts. The proposal draws on the Transparency and Openness Promotion (TOP) data-sharing guidelines, the *Mata v. Avianca* legal precedent, and USPTO prior-art submission practices. It addresses fabrication directly, because an author cannot deposit the full text of a source that does not exist. The Source Provenance Ledger differs from Glynn's proposal in two respects: it is a private record maintained by the author rather than a mandatory submission component, and it extends beyond full-text deposit to include status checks, archive timestamps, claim-level support mapping, and a recheck trigger for high-risk sources.

## 9. The Source Provenance Ledger

The Source Provenance Ledger is a private evidentiary record that an author creates and maintains at the point of citation. It documents what was checked, when, what the source said, whether it was live, whether it had been corrected or retracted, and where the preserved copy resides. The record remains under the author's control, comparable to research notes or source files. It is never submitted as a default part of manuscript submission.

### 9.1 Proposed Fields

The ledger operates at three tiers of depth, selected by the author based on the stakes of the claim the citation supports. The author decides which tier to apply to each citation based on the author's own risk assessment.

**Tier 1: Basic Source Check (all references).** Source title, author or organization, DOI or URL, date accessed, and existence confirmation (the URL resolved to a live page with matching title and author). Estimated time: under one minute per citation.

**Tier 2: Claim Support Record (references supporting factual, statistical, medical, policy, legal, or accusatory claims).** All Tier 1 fields plus the exact claim the source is cited to support, the page or section where the supporting evidence appears, and a claim-

support classification (direct, indirect, background only, or does not support). The claim-support field addresses a risk the UK Research Integrity Office identifies directly: AI tools "may also provide real looking references that do not support the claim being made," making existence verification alone insufficient (UKRIO, n.d.). Estimated time: two to three minutes per citation.

**Tier 3: Full Source Custody Record (high-risk claims where challenge is anticipated or consequences of error are significant).** All Tier 2 fields plus an archive link (Internet Archive or Perma.cc where available), retraction or correction status at time of access documented through Crossmark or Retraction Watch, a later review date for time-sensitive sources, and a one-sentence author diligence note explaining why the source was considered reliable. Estimated time: three to five minutes per citation.

For a manuscript with 60 references, of which 20 support factual claims at Tier 2 and 5 carry high-risk claims at Tier 3, the estimated total documentation time is approximately 60 to 90 minutes. This is comparable to the time researchers currently spend formatting reference lists and often less than the time spent writing the manuscript itself. The cost is real and should be weighed against the cost of having no evidence to produce when a citation is challenged.

No external party decides which citations receive which tier. The author makes that judgment based on the nature of the claim, the stability of the source, and the author's assessment of the likelihood that the citation will be challenged. An opinion editorial may warrant no ledger at all. A blog post may warrant Tier 1 for a few key citations. A book manuscript approaching final publication may warrant Tier 3 for every factual claim. The decision belongs to the author.

### 9.2 Private Record, Producible on Challenge

The ledger is not submitted with the manuscript. It is maintained by the author as part of their research governance practice and produced at the author's discretion when a citation is challenged or the work's credibility is questioned. If the author chooses not to produce the record, that is the author's right. If the author declines to produce the record, that choice should be treated as discretionary unless a publisher, institution, or legal process requires production.

If a challenge proceeds before a private institution (journal, publisher, university review board), the author can decide whether to share the ledger, what portions to share, and under what conditions. The author should have the right to require a non-disclosure agreement or request that the ledger remain confidential beyond the review board to protect intellectual property, methodology, or source relationships. A reviewing body can request the full ledger for all disputed citations. The author can decline, but declining after being asked is itself information the adjudicator may weigh.

If allegations are brought before a legal entity, production may be compelled through legal process, subject to the same protections that govern any evidentiary production.

Parts of the record may be unproducible regardless of the author's willingness, because of HIPAA protections (where applicable), privacy regulations, anonymous data-collection protocols, or other legal constraints on the underlying source material. The ledger's design accommodates this: the diligence note can document that a source was verified under conditions that prevent full disclosure without specifying the protected content.

### 9.3 Three Channels of Adoption

The ledger's adoption is voluntary for authors. Three independent channels create incentives for its use, and each operates at a different point in the publication lifecycle.

**Author prerogative.** The author decides which outputs warrant governed documentation based on their own assessment of stakes. A personal blog carries no expectation of a ledger. A research paper approaching final publication warrants thorough documentation because the author bears the consequence of having nothing to produce if challenged. The cost of maintaining the record is borne by the person who benefits from having it. The author can choose not to maintain a ledger for any piece of work, for any reason, without automatic negative inference. The reason can be as simple as "this was an opinion editorial" or "this was a working paper and I did not anticipate this level of scrutiny."

**Publisher requirement.** Journals and platforms can require that submitting authors maintain a Source Provenance Ledger for any work they submit, producible on request if a citation is challenged. This is analogous to existing submission requirements for data availability statements, ethics declarations, and conflict-of-interest disclosures. The journal does not review the ledger at submission. It requires the author to attest that one exists and can be produced. A journal can calibrate: require attestation for research articles, recommend it for reviews, waive it for editorials. This requirement helps reduce risk to the publishing entity by ensuring documentation exists before a dispute arises. The concentration of academic publishing makes this channel practical: the twelve largest publishers account for approximately two-thirds of all scholarly output, so adoption by even a fraction of them would cover a substantial share of the published record (Sousa et al., 2026).

**Insurance condition.** Errors-and-omissions or directors-and-officers insurance policies covering publication-related claims can condition coverage on the maintenance of source custody records. If a citation dispute produces a claim and the policyholder maintained a ledger, the claim could be covered. If the policyholder did not maintain one, coverage could be denied or limited. This creates a market-driven incentive for adoption without any government mandate, regulatory requirement, or ethics-body directive. The insurer does not mandate verification. The insurer prices the risk. Authors and institutions that maintain records are insurable at lower cost. Those that do not bear the cost of disputed claims alone. This pattern is already established in medical malpractice, legal malpractice, and financial services errors-and-omissions coverage, where record-keeping affects both insurability and claim outcomes.

### 9.4 What the Ledger Produces at Each Fault Level

At Levels 1 and 2 (fabrication and failure to verify), the inability to produce any ledger documentation is consistent with the absence of verification. An author who never checked cannot produce evidence of checking. The mechanism is asymmetric in the right direction: it shields the diligent who kept a record and exposes the careless who did not. A careless author cannot easily recreate contemporaneous third-party timestamps or Crossmark query logs retroactively.

At Level 3 (negligent verification), a ledger entry showing that the author accessed the source and recorded the passage they relied on, but misinterpreted its meaning, is evidence of good-faith engagement. Correction is warranted. The evidentiary record distinguishes misinterpretation from non-engagement.

At Level 4 (downstream contamination), a ledger entry showing that the source was not retracted or flagged at time of access, documented through a Crossmark or Retraction Watch check, is evidence that the author verified and the source status changed after verification. The fault lies upstream.

At Level 5 (source decay), a ledger entry with an archive link or local copy is evidence that the source existed and supported the claim. The medium failed. The author did not.

The ledger does not immunize an author. It creates reviewable evidence of what the author checked, when the author checked it, and whether the author's reliance was reasonable under the conditions available at the time.

## 10. Limitations, Dissent, and Validation Roadmap

### 10.1 Limitations

The fault ladder has not been adopted by any publisher, journal, ethics body, or repository. It is a proposed framework supported by an evidence base, not an established standard. The Source Provenance Ledger has not been tested in a live submission workflow. Its field specifications are theoretical, derived from the verified evidence about what verification tools exist and what information they produce, but this paper does not document a live author workflow maintained under these specifications and no editor has yet adjudicated a contested citation using one.

The evidence base for AI-specific citation cascading remains preliminary. The GhostCite study (Xu et al., 2026, preprint) found the same fabricated citation appearing across up to 16 separate papers, which is the one of the strongest empirical observations of AI-driven citation propagation identified for this paper. No published study has yet traced the full chain from a single hallucinated citation through every downstream paper that incorporated it. The retraction-persistence evidence (Hsiao & Schneider, 2021; Bakker et al., 2024; De Oliveira Andrade, 2022) documents the mechanism at scale for retracted papers, and the inference that AI-fabricated citations would propagate through the same channel is supported by the structural parallels, but direct measurement of the full AI-to-downstream propagation rate remains an empirical gap that future work should address.

The field-variation data (Tramèr, 2025; Xu et al., 2026; Topaz et al., 2026) shows contamination rates that differ by more than an order of magnitude across disciplines. This variation means the fault ladder's thresholds may need to be calibrated differently for fields with high baseline contamination rates versus fields where fabricated citations are rare. The paper does not propose field-specific calibrations and acknowledges that doing so would require discipline-level empirical work beyond the current scope.

The Lancet study (Topaz et al., 2026) anchors the paper's scale claims. It is a peer-reviewed research letter, a lighter editorial form than a full research article. Published methodological commentary has questioned how the AI-assisted detection pipeline was validated and whether the audit distinguishes scientifically material fabricated citations from immaterial ones (Naddaf, 2026; Hosseini & Resnik, 2026 [PROVISIONAL]). This paper cites the Topaz figures as the best available quantitative evidence while acknowledging the live methodological debate.

The multi-platform parallel research methodology used to assemble this paper's evidence base is described in Appendix C. The methodology dispatched identical research prompts across multiple AI platforms, followed by convergence analysis and independent source verification. It is a practitioner methodology without formal peer-reviewed validation. Agreement across platforms may reflect shared training data rather than independent assessment, and the methodology should be understood as a literature-assembly tool rather than an independent source of evidence. The paper's empirical weight rests on the human-authored studies it cites (Topaz, Xu, Hsiao, Schneider, Klein, Zittrain, Bakker, Avenell, Van der Vet, Nijveen, Glynn), not on the cross-platform convergence.

The fault ladder grades author conduct and does not include a materiality axis (whether the failed citation affected the paper's conclusions). Hosseini and Resnik have argued that the impact of a hallucinated citation on the paper's scientific findings should factor into the response. This paper takes a different position: if an author can produce evidence of diligence at the time of use, the author acted in good faith regardless of the citation's weight in the argument. Materiality is relevant to the editorial response (how urgently to correct, whether to issue an expression of concern) but is a separate question from author culpability. The editor evaluates the damage to the record. The ledger evaluates the author's conduct. These are different assessments with different owners. Whether to add a materiality axis to the fault ladder itself is a design question for future work.

### 10.2 The Enforcement Simplicity Objection

The strongest counter-argument to the fault ladder is administrative. A uniform enforcement standard is administratively simple: it requires only checking whether the citation is valid. If it is not valid, the author is responsible. The evaluation ends there. A fault-based system requires someone to evaluate what the author knew, what the author did, and whether the author's verification was reasonable under the circumstances. That evaluation is subjective, resource-intensive, and invites strategic behavior by authors who will claim the lowest fault level available regardless of what actually happened.

A strict-liability defender would add that any defense introduced into a system that currently under-enforces lowers the expected cost of carelessness further. The measured

problem is that almost no one is corrected or sanctioned (98.4% of flagged papers uncorrected, 76.7% of reviewers not checking references). Adding a defense to a regime that barely enforces is, by this argument, a solution pointed at the wrong failure.

Three responses are available, and all have limits.

First, the under-enforcement datum is the motivation for the paper, not a refutation. Manual enforcement failed. That is why enforcement is now being automated. The automated-enforcement wave (citation-verification pipelines, AI-detection products, platform-level sanctions triggered by algorithmic flags) will reach authors who were previously untouched. When it does, it will produce false positives, and the authors falsely flagged will need a mechanism to produce evidence of diligence. The ledger is that mechanism. The paper does not propose a defense against enforcement that exists. It proposes a rebuttal mechanism for enforcement systems that are emerging.

Second, the ledger shifts the evidentiary burden to the author. Under the current system, an adjudicator must determine whether the author checked. Under the proposed system, the author produces the record or does not. If the author produces a complete ledger entry with external verification (Internet Archive timestamps, Crossmark logs, Retraction Watch queries), adjudication is faster because the evidence is documented and independently verifiable. If the author cannot produce any record, the absence is itself informative and adjudication can proceed under the current standard. The ledger does not add a five-level hearing to every case. It adds a single checkpoint: can the author produce evidence of diligence, yes or no?

Third, the three-channel adoption model (Section 9.3) does not require every platform or journal to implement the full fault ladder. A preprint server operating at high volume may reasonably apply a simpler standard (clear evidence of unchecked AI output triggers a ban) while a journal publishing final peer-reviewed work may apply the full graduated framework. The fault ladder is a governance architecture, not a universal mandate.

The adjudication cost of a fault-based system remains an empirical question. Whether the ledger reduces or increases the time to resolution for disputed citations can only be measured through the pilot testing proposed in Section 10.4. The paper does not claim the fault ladder is administratively simpler than the current system. It claims the current system produces outcomes that make verification effort invisible, and that a system which makes effort legible is worth testing.

### 10.3 Dissent to Preserve

Three substantive tensions exist in the evidence base and should be preserved rather than resolved.

The deterrence-versus-documentation tension is the most consequential. Platforms need strong deterrents against fabricated citations. A documented diligence record, if recognized, could be misused by authors who game the documentation process. The counter-position is that deterrence without a documentation path makes diligence invisible, which removes institutional incentive to verify. The mechanism is asymmetric:

only the author who maintained a record can produce one, and contemporaneous external timestamps from the Internet Archive or Crossmark cannot be created retroactively. A signed attestation alone is insufficient. Selective disclosure (producing a curated subset of clean entries while withholding incomplete ones) is a real opportunity for strategic misuse that the framework must acknowledge. A reviewing body that requests a full ledger for all disputed citations, with the author's right to decline noted, creates a signal that partial production cannot obscure.

The gap between COPE's graduated philosophy and arXiv's blunt enforcement instrument remains unresolved. This paper argues the fault ladder bridges it, but neither COPE nor arXiv has endorsed a graduated approach with an evidentiary record, and individual publishers may continue to apply their own standards regardless of what this paper proposes. The field's response is an empirical question.

The balance between false-positive protection and enforcement rigor is genuinely unsettled. Automated detection will reduce the volume of uncaught fabrications, but it will also produce wrongful flags. Whether the marginal harm of a wrongful flag outweighs the marginal benefit of catching a fabrication depends on the detection tool's accuracy, the consequences of each error type, and the availability of a rebuttal mechanism. The paper argues the rebuttal mechanism is necessary regardless of where the balance falls.

### 10.4 Validation Roadmap

Validation should proceed in three stages.

**Stage 1: Pilot testing with a single journal or preprint server.** Partner with one journal or repository to request that submitting authors maintain a Source Provenance Ledger, producible on request. Measure: how many authors maintain ledger entries, how complete they are, what fraction of citations can be verified using the ledger when challenged, whether editors find the ledger useful in adjudicating citation disputes, and the time cost per citation at each tier.

**Stage 2: Retrospective testing against known cases.** Apply the fault ladder retroactively to a sample of retracted papers and papers flagged for citation errors. Classify each case by fault level using the decision tree in Section 7.7 and determine whether the author could have produced a ledger entry that would have changed the adjudication outcome. Measure: the proportion of cases where a ledger entry would have provided meaningful evidence, the proportion where it would not, and inter-rater reliability when three independent classifiers apply the decision tree to the same cases.

**Stage 3: Integration testing with citation managers.** Build a prototype citation-manager plugin that assists ledger entry creation at the point of citation. Test it with a cohort of researchers across multiple disciplines. Measure: adoption rates, time cost per citation at each tier, completeness of generated entries, and researcher perception of the tool's value. This plugin is an author's tool, not a journal requirement. It assists voluntary record-keeping.

## 11. Conclusion

Manual enforcement of citation integrity has failed. The evidence presented here points in one direction: 98.4% of papers flagged for fabricated citations remain uncorrected, 76.7% of peer reviewers do not check references, and 41.5% of researchers copy-paste citations without verifying them. The human verification chain is broken at every level, and the field's response is to automate enforcement through citation-verification pipelines, AI-detection tools, and platform-level sanctions. That automation will catch fabrications that humans missed. It will also flag authors who did nothing wrong.

The current system has no mechanism for an author to produce evidence of diligence when a citation is challenged or flagged. An author who verified every source, ran status checks, and archived the evidence occupies the same position as an author who never opened a single reference, because no framework records or produces the difference. Verification effort is invisible. When effort is invisible, it cannot be rewarded, it cannot be distinguished from its absence, and the incentive to invest in it collapses to personal conscience. The behavioral data shows that personal conscience is not sufficient.

The fault ladder separates source failures into five levels calibrated to the author's conduct: fabrication, failure to verify, negligent verification, downstream contamination, and source decay. It bridges the gap between COPE's graduated philosophy (responsibility established as principle, sanctions reserved for misconduct, education as the default response) and arXiv's blunt enforcement instrument (a one-year ban for "incontrovertible evidence" of unchecked output). The ladder preserves strong enforcement at Levels 1 and 2, where it belongs, while honoring COPE's anti-ban guidance at Levels 3 through 5, where the author's conduct does not warrant blanket sanction.

The Source Provenance Ledger makes the distinction between these levels producible. It is a private evidentiary record maintained by the author, documented at the point of citation, and produced when the author's work is challenged. It follows the precedent by which attorneys maintain work product, physicians maintain clinical records, journalists maintain source files, and financial professionals maintain audit trails: private documentation, produced on challenge, with the ability to produce records affecting the outcome of disputes.

Adoption is distributed across three channels, each with its own incentive structure. The author decides which outputs warrant documentation based on stakes. Publishers can require that submitting authors maintain a ledger producible on request. Insurance carriers can condition coverage on source custody records, following the pattern by which professional liability insurance has historically motivated record-keeping in law, medicine, and financial services. The insurer does not mandate verification. The insurer prices the risk.

The core function of the ledger is legibility. It makes verification effort visible. An author who documented can produce the record. An author who did not cannot fabricate it retroactively. Defense against false accusation, deterrence against carelessness, and fair adjudication of disputed citations all follow from making effort visible. The mechanism is asymmetric in the right direction, and market incentives can support adoption without requiring a mandate.

## References


Ansari, S. (2026). Compound deception in elite peer review: A failure mode taxonomy of 100 fabricated citations at NeurIPS 2025. *arXiv preprint arXiv:2602.05930*. https://arxiv.org/abs/2602.05930

Avenell, A., Stewart, F., Grey, A., Gamble, G., & Bolland, M. (2019). An investigation into the impact and implications of published papers from retracted research: Systematic search of affected literature. *BMJ Open*, *9*, e031909. https://doi.org/10.1136/bmjopen-2019-031909

Bakker, C., Boughton, S., Faggion, C. M., Fanelli, D., Kaiser, K., & Schneider, J. (2024). Reducing the residue of retractions in evidence synthesis: Ways to minimise inappropriate citation and use of retracted data. *BMJ Evidence-Based Medicine*, *29*(2), 121–126. https://doi.org/10.1136/bmjebm-2022-111921

Committee on Publication Ethics. (2023, updated 2025, February 24). Authorship and AI tools: COPE position statement. https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools

Committee on Publication Ethics. (2026, April 22). Glossary. https://publicationethics.org/glossary

De Oliveira Andrade, R. (2022, December). Retracted scientific articles continue to be cited by other scientists as reliable sources. *Pesquisa FAPESP*, (322). https://revistapesquisa.fapesp.br/en/retracted-scientific-articles-continue-to-be-cited-by-other-scientists-as-reliable-sources/

Dulin, K., & Ziegler, A. (2017). Scaling up Perma.cc: Ensuring the integrity of the digital scholarly record. *D-Lib Magazine*, *23*(5/6). https://www.dlib.org/dlib/may17/dulin/05dulin.html

Federal Rules of Civil Procedure, Rule 11(b). https://www.law.cornell.edu/rules/frcp/rule_11

Glynn, A. (2025). Guarding against artificial intelligence-hallucinated citations: The case for full-text reference deposit. *European Science Editing*, *51*. https://doi.org/10.3897/ese.2025.e153973

Hsiao, T.-K., & Schneider, J. (2021). Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. *Quantitative Science Studies*, *2*(4), 1144–1169. https://doi.org/10.1162/qss_a_00155

Hosseini, M., & Resnik, D. (2026). [PROVISIONAL] On distinguishing scientifically material hallucinated citations from immaterial ones. Referenced via coverage of the Topaz et al. audit; primary source not retrieved in full for this paper.

Jibril, A. B. (2025). False authorship: An explorative case study. *PubMed Central*. https://pmc.ncbi.nlm.nih.gov/articles/PMC12107892/

Jones, S. M., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R., & Grover, C. (2016). Scholarly context adrift: Three out of four URI references lead to changed content. *PLOS ONE*, *11*(12), e0167475. https://doi.org/10.1371/journal.pone.0167475

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly context not found: One in five articles suffers from reference rot. *PLOS ONE*, *9*(12), e115253. https://doi.org/10.1371/journal.pone.0115253

Lin, J., Shan, R., Zhu, J., Xi, Y., Yu, Y., & Zhang, W. (2025). Stop DDoS attacking the research community with AI-generated survey papers. *arXiv preprint arXiv:2510.09686*. https://arxiv.org/abs/2510.09686

*Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443 (S.D.N.Y. 2023). https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/

Chawla, D. S. (2026, May 19). Researchers who use hallucinated references to face arXiv ban. *Nature*. https://www.nature.com/articles/d41586-026-01595-5

Naddaf, M. (2026). Surge in fake citations uncovered by audit of 2.5 million biomedical-science papers. *Nature*. https://www.nature.com/articles/d41586-026-00748-w

Parker, L., Boughton, S., Bero, L., & Byrne, J. A. (2024). Paper mill challenges: Past, present, and future. *Journal of Clinical Epidemiology*, *176*, Article 111549. https://doi.org/10.1016/j.jclinepi.2024.111549

Puglisi, B. C. (2025). *Governing AI: When capability exceeds control*. ISBN 9798349677687.

Research Information. (2026, May). arXiv imposes one-year ban for unchecked AI-generated content. https://www.researchinformation.info/news/arxiv-imposes-one-year-ban-for-unchecked-ai-generated-content/

Sousa, S., Rowcliffe, N., Iorio, B., & Perchyk, T. (2026). Managing the risks of generative AI in academic publishing. *Global Institute for Economy and Finance*.

The Scientist. (2026, May). One in 277 biomedical papers carry fake references. https://www.the-scientist.com/one-in-277-biomedical-papers-carry-fake-references-74480

Topaz, M., Roguin, N., Gupta, P., Zhang, Z., & Peltonen, L.-M. (2026). Fabricated citations: An audit across 2.5 million biomedical papers. *The Lancet*, *407*(10541), 1779–1781. https://doi.org/10.1016/S0140-6736(26)00603-3

Tramèr, F. (2025, August 3). Trends in LLM-generated citations on arXiv. *SPY Lab Blog*, ETH Zurich. https://spylab.ai/blog/hallucinations/

UK Research Integrity Office. (n.d.). AI in research. https://ukrio.org/ukrio-resources/ai-in-research/

Van der Vet, P. E., & Nijveen, H. (2016). Propagation of errors in citation networks: A study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. *Research Integrity and Peer Review*, *1*(3). https://doi.org/10.1186/s41073-016-0008-5

Xu, Z., Qiu, Y., Sun, L., Miao, F., Wu, F., Wang, X., Li, X., Lu, H., Zhang, Z., Hu, Y., Li, J., Luo, J., Zhang, F., Luo, R., Liu, X., Li, Y., & Liu, J. (2026). GhostCite: A large-scale analysis of citation validity in the age of large language models. *arXiv preprint arXiv:2602.06718*. https://arxiv.org/abs/2602.06718

Zittrain, J., Albert, K., & Lessig, L. (2014). Perma: Scoping and addressing the problem of link and reference rot in legal citations. *Harvard Law Review Forum*, *127*, 176–199. https://harvardlawreview.org/forum/vol-127/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations/

## Supplementary Sources

Inside Higher Ed. (2026, May 22). Ban on authors who submit AI content “welcome but unenforceable.” https://www.insidehighered.com/news/faculty/books-publishing/2026/05/22/ban-authors-who-submit-ai-content-welcome-unenforceable

TechCrunch. (2026, May 16). Research repository ArXiv will ban authors for a year if they let AI do all the work. https://techcrunch.com/2026/05/16/research-repository-arxiv-will-ban-authors-for-a-year-if-they-let-ai-do-all-the-work/

## Appendix A: Source Provenance Ledger Field Template and Verification Gate

The following fields constitute the minimum viable record for a single citation entry, organized by tier. The author selects the appropriate tier based on the stakes of the claim the citation supports.

### Tier 1: Basic Source Check (all references)

| Field | Description |
|---|---|
| Source title | Full title as it appears on the source page |
| Author or organization | Author names or institutional publisher |
| DOI, URL, or persistent identifier | Persistent identifier or full URL accessed |
| Date accessed | Date the citing author opened and reviewed the source |
| Existence confirmed | The URL resolved to a live page with matching title and author |

### Tier 2: Claim Support Record (references supporting factual, statistical, medical, policy, legal, or accusatory claims)

| Field | Description |
|---|---|
| All Tier 1 fields | As above |
| Exact claim used | The specific assertion the citation is intended to support, stated in one sentence |
| Page, paragraph, section, or timestamp | Location within the source where the supporting evidence appears |
| Claim support classification | Direct support, indirect support, background only, or does not support |
| Source type | Primary research, secondary source, commentary, data, opinion, or news reporting |

### Tier 3: Full Source Custody Record (high-risk claims where challenge is anticipated)

| Field | Description |
|---|---|
| All Tier 1 and Tier 2 fields | As above |
| Archive link | Internet Archive link, Perma.cc link (where institutional access exists), or local file hash |
| Retraction or correction status at time of access | Result of Crossmark, Retraction Watch, or publisher status check |
| Status check method | Which tool or method was used to verify current status |
| Later review date | Scheduled recheck date for high-risk or time-sensitive sources |
| Later status change | Any correction, retraction, or removal discovered after publication |
| Author diligence note | One sentence explaining why the source was considered reliable at time of use |

### Verification Gate (at time of citation)

At Tier 1: Open the source URL. Confirm the page loads with content. Confirm the title and author match. Record the access date.

At Tier 2: Complete all Tier 1 steps. Read the relevant passage. Record the specific claim and its location in the source. Classify the claim support.

At Tier 3: Complete all Tier 1 and Tier 2 steps. Check the retraction and correction status through Crossmark, Retraction Watch, or the publisher's correction page. Create an archive snapshot through the Internet Archive or Perma.cc (where institutional access exists), or capture a local PDF. Write a one-sentence diligence note. If the source cannot be opened, does not match the metadata, does not support the claim, or has been retracted or corrected, exclude it from the reference list or document the reason for its inclusion.

## Appendix B: Worked Lifecycle Example

The following example traces a single citation through the four post-publication failure modes and shows what the Source Provenance Ledger would contain at each stage.

**Initial state.** An author cites a web-hosted research report in a manuscript about workplace AI adoption. At the time of citation, the author opens the URL, confirms the page loads, reads the relevant section, records the specific claim (that 37% of surveyed organizations had adopted AI-assisted hiring tools by 2025), creates an Internet Archive snapshot, checks Retraction Watch (not applicable for industry reports), and writes a

diligence note: “Source is a published report from a named industry research organization, accessed directly, claim appears on page 14 of the PDF, archived via Internet Archive.” The ledger entry is maintained in the author’s private records.

**Scenario A: Source decay (link rot).** Six months after publication, the research organization restructures its website. The original URL returns a 404 error. If the citation is challenged, the author produces the Internet Archive snapshot from the ledger, preserving the page as it existed at time of citation. The ledger entry shows the source existed and supported the claim. No correction is needed beyond adding an editorial note directing readers to the archived version. Fault level: 5 (source decay). No sanction warranted.

**Scenario B: Content drift.** The research organization updates the report with new survey data. The URL still resolves, but the 37% figure has been revised to 42%. If challenged, the author produces the archive snapshot preserving the original version. The ledger entry documents which version was cited and when. A correction note can acknowledge the update while preserving the accuracy of the original citation. Fault level: 5 (source decay variant). No sanction warranted.

**Scenario C: Downstream contamination.** A year after publication, the original research report is withdrawn because the survey methodology was found to contain a sampling error. The author’s Retraction Watch check at time of use showed no flag. The ledger entry documents that the author checked and the source was clean at that moment. Upon learning of the withdrawal, the author issues a correction. Fault level: 4 (downstream contamination). No sanction warranted. The author can produce evidence of good faith.

**Scenario D: Exposed fabrication.** An investigation reveals that the original research report was generated by an AI tool and the research organization does not exist. The author’s ledger entry shows the page loaded, the content appeared coherent, the organization’s name appeared legitimate, and no red flags were visible at time of access. The author’s diligence was genuine but the source was a sophisticated fabrication. The fault lies with whoever created and hosted the fake report. The citing author’s ledger entry provides evidence that a reasonable verification effort was made. Depending on the sophistication of the fabrication and the ease of detection, this falls between fault levels 3 and 4. The ledger does not immunize the author. It provides the evidentiary basis for distinguishing this author from one who never checked at all.

## Appendix C: Author Background, Methodology, and Related Work

The practitioner methodology that produced this work began with the Factics framework (Facts plus Tactics with measurable outcomes), developed in 2012 as a content evaluation methodology and extended across multiple publications including *Digital Factics X* (Kirkus-reviewed) and *Governing AI: When Capability Exceeds Control* (Puglisi, 2025). The operational context includes twelve years in law enforcement at a major metropolitan airport, which established the checkpoint-verification discipline that informs the governance architecture throughout this work.

The paper's evidence base was assembled through a multi-platform parallel research methodology (HAIA-CAIPR), which structures the dispatch of identical research prompts across multiple AI platforms followed by convergence analysis and independent human-led source verification. The methodology is a literature-assembly tool. Agreement across platforms may reflect shared training data rather than independent assessment, and it should not be treated as independent evidence. The paper's empirical weight rests on the human-authored studies it cites. The cited sources selected for this paper were verified by opening the primary URL, confirming title and author metadata, and confirming that the content supports the claim assigned to it. Citation verification problems documented in this paper were encountered through multi-platform practice during 2025 and 2026, with findings preserved under the HAIA-CARCS (Compliance Accountability Record and Case Study) documentation protocol.

All governance frameworks referenced in this paper (Checkpoint-Based Governance, GOPEL, HAIA-CAIPR, HAIA-CARCS) are published open-source under Creative Commons at github.com/basilpuglisi/HAIA. The AI Provider Plurality Congressional Package was submitted to the 119th Congress in February 2026 (Puglisi, 2026). None of this work is affiliated with any corporation, funded by any investor, or peer-reviewed by a formal journal review process.

## Appendix D: Source Verification Screenshots

The following pages contain browser screenshots of every cited source in this paper, captured on June 2, 2026. Each screenshot confirms the source URL loads, the title and author metadata match the reference list, and the page contains content relevant to the claim assigned to it. Three sources are not captured as screenshots: Klein et al. (2014) and Parker et al. (2024) were verified via DOI resolution and direct page access during the research session; Naddaf (2026) in Nature returned a 502 server error at capture time but was verified live earlier in the session and its content is independently confirmed through the Research Information article covering the same enforcement action.

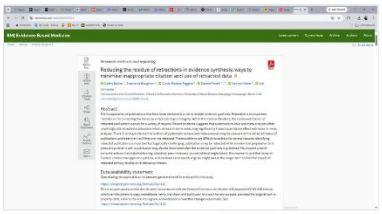
This appendix practices what the paper proposes. It is a source custody record for the paper's own citations.

## Appendix D: Source Verification Screenshots (Page 1 of 5)

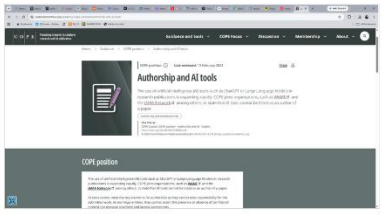
Captured June 2, 2026. Each screenshot confirms the cited source loads and matches paper claims.

**Ansari (2026)**
arXiv:2602.05930

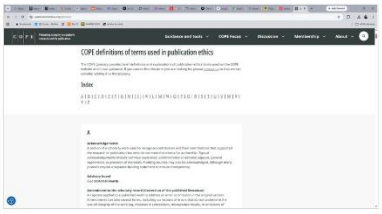
**Avenell et al. (2019)**
BMJ Open



**Bakker et al. (2024)**
BMJ Evidence-Based Medicine




**COPE Position Statement (2023)**
publicationethics.org



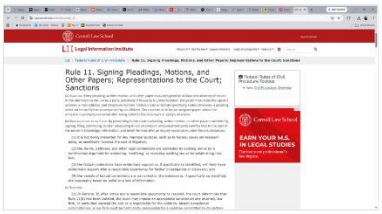

**COPE Glossary (2026)**
publicationethics.org/glossary


**De Oliveira Andrade (2022)**
Pesquisa FAPESP

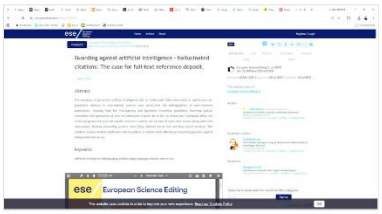
## Appendix D: Source Verification Screenshots (Page 2 of 5)

Captured June 2, 2026. Each screenshot confirms the cited source loads and matches paper claims.

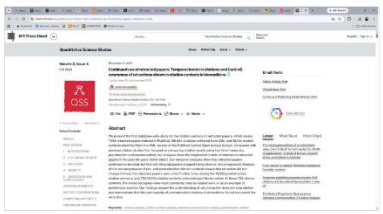
**Dulin & Ziegler (2017)**
D-Lib Magazine




**FRCP Rule 11(b)**
law.cornell.edu

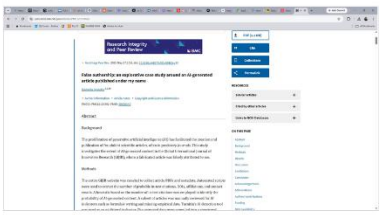



**Glynn (2025)**
European Science Editing



**Hsiao & Schneider (2021)**
Quantitative Science Studies

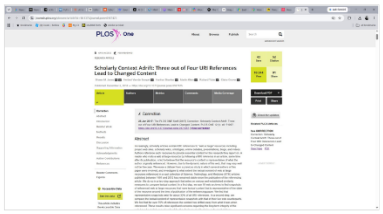

**Hosseini & Resnik (2026)**
Accountability in Research



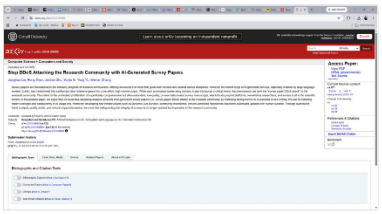

**Jibril (2025)**
Research Integrity & Peer Review

## Appendix D: Source Verification Screenshots (Page 3 of 5)

Captured June 2, 2026. Each screenshot confirms the cited source loads and matches paper claims.

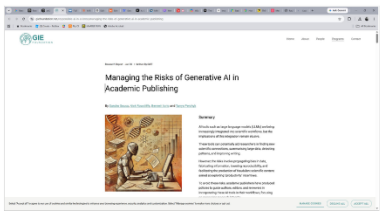



**Jones et al. (2016)**
PLOS ONE

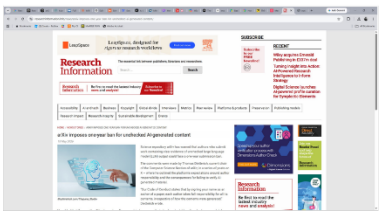



**Lin et al. (2025)**
arXiv:2510.09686

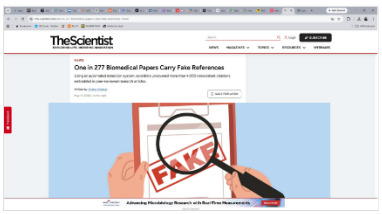

**Mata v. Avianca (2023)**
Justia / S.D.N.Y.




**Sousa et al. (2026)**
GIEF Foundation



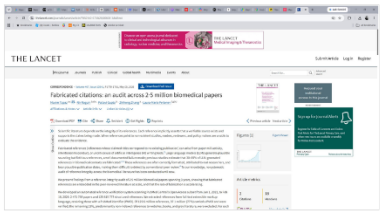

**Research Information (2026)**
researchinformation.info



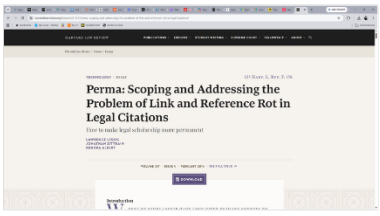

**The Scientist (2026)**
the-scientist.com

## Appendix D: Source Verification Screenshots (Page 4 of 5)

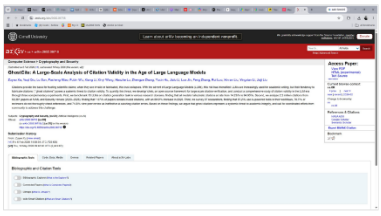
Captured June 2, 2026. Each screenshot confirms the cited source loads and matches paper claims.




**Topaz et al. (2026)**
The Lancet

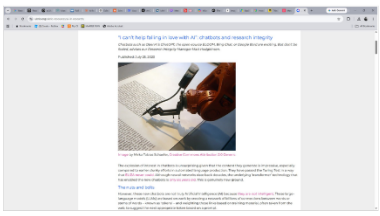



**Zittrain et al. (2014)**
Harvard Law Review Forum

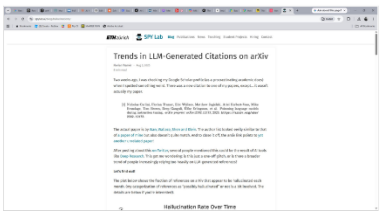



**Xu et al. / GhostCite (2026)**
arXiv:2602.06718


**Van der Vet & Nijveen (2016)**
Research Integrity & Peer Review



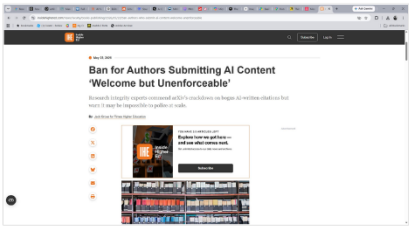

**UKRIO (n.d.)**
ukrio.org



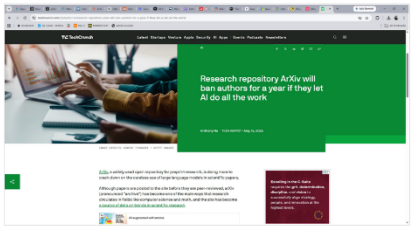

**Tramer (2025) / SPY Lab**
spylab.ai / ETH Zurich

## Appendix D: Source Verification Screenshots (Page 5 of 5)

Captured June 2, 2026. Each screenshot confirms the cited source loads and matches paper claims.




**Inside Higher Ed (2026)**
insidehighered.com




**TechCrunch (2026)**
techcrunch.com

Note: Chawla/Nature (2026) DOI 10.1038/d41586-026-01595-5 returned 502 at capture time.
Content independently confirmed via Research Information (screenshot above) and prior session verification.
Klein et al. (2014), Naddaf (2026), Parker et al. (2024): verified via DOI resolution and web_fetch; no screenshot captured.