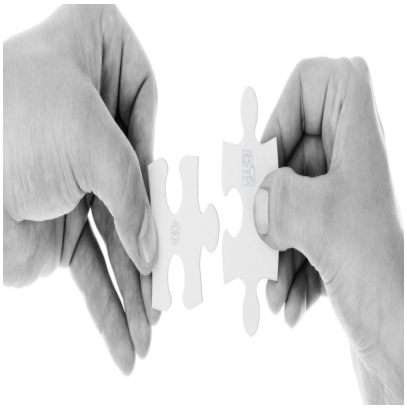

Description

Every system the scholarly publishing industry has built in the last three years to fight AI misuse shares one assumption: that the thing worth catching is deception. The STM Integrity Hub screens for tortured phrases, duplicate submissions, and the textual fingerprints of paper mills. COPE's 2025 retraction guidelines now name the undisclosed use of artificial intelligence in the same breath as fraud, identity theft, and fictitious authorship; a unified category of *misrepresentation*. Detection tools are tuned to find the manuscript that's lying about what it is.

This is good infrastructure for the crisis we already had. It is close to futile for the one quietly building underneath it.

The numbers make the distinction concrete. An analysis of 5,114 journals and more than 5.2 million papers and found that roughly 70% of journals now have some form of AI disclosure policy. Of 75,000 papers published since 2023 with full-text analysis available, only about 0.1% explicitly disclosed AI use. Separately, when the American Association for Cancer Research ran AI detectors against 7,177 manuscripts submitted across six months in 2024, 36% of abstracts showed signs of AI-generated text. Only 9% of authors had disclosed any AI use when asked directly in the submission process.

Read those two studies side by side and a question forms that fraud-detection tools were never built to answer: if essentially everyone is quietly using AI and essentially no one is disclosing it, is this fraud? Or is it something the field hasn't actually defined yet?

Two Different Problems Wearing the Same Coat

It's worth being precise about what "undisclosed AI use" is actually describing, because most industry conversation treats it as one thing. A paper mill that fabricates data and runs it through a language model to evade plagiarism detectors is committing fraud. The deception is the point; the AI is just a faster brush.

A working scientist who runs a draft paragraph through an LLM to tighten the prose, doesn't think to mention it because their journal's policy never defined "substantial use," and submits a paper whose findings are completely sound. That is not fraud. That's a researcher operating inside a disclosure framework that was written before anyone agreed on what needed disclosing.

These two scenarios produce an identical signal to a text detector: AI-generated language, no disclosure statement. They are not remotely the same integrity event. One is concealment. The other is ambiguity about what concealment even means in a workflow where AI assistance is now so routine that one analyst, commenting on the PNAS findings, noted that publishers haven't even agreed on whether something as basic as light editing requires disclosure at all. When the category boundary itself is undefined, "non-compliance" stops being a meaningful description of what's happening. It's not that authors are choosing not to follow the rule. It's that the rule doesn't yet specify a rule.

Provenance, Not Authorship, is the Thing Breaking

The "tool-versus-author" positioning questions if an AI system can hold the position of an author and who is answerable when something goes wrong. This is a real and serious debate, and one I've written about elsewhere in the context of autonomous AI research systems. But it's a different fault line from the one the PNAS data is exposing.

Provenance ambiguity isn't about whether AI wrote the paper. It's about the fact that, for a clear majority of papers being published right now, nobody traces what AI touched, how much, or where. Not because anyone is hiding it, but because the infrastructure for tracking contribution at that level of granularity doesn't really exist yet. A researcher who used an LLM to summarize background literature, another who used it to generate a first-pass methods description, and another who used it to polish grammar in a final draft are doing three different things with three different implications for the reliability of the resulting paper. Right now, all three would show up, if they showed up at all, as an identical checkbox: "AI was used."

That's not a disclosure problem you solve by demanding more disclosure. It's a *resolution* problem. The category is too coarse to carry the information the field actually needs.

This matter because provenance, in research integrity terms, has always meant something specific: a documented, traceable chain showing where a claim, a dataset, or a method came from, and who is accountable for each step. Citation systems exist to make that chain visible. Methods sections exist to make it reproducible. Authorship statements exist to make it accountable. AI use is the first major workflow input in decades that can sit inside all three of those structures; woven into the prose, the methods, even the analysis, while leaving none of the usual traces that let an editor or reader reconstruct what actually happened.

Why Detection Infrastructure Can't Close This Gap

It's tempting to read the AACR numbers, 36% detected, 9% disclosed, as evidence that publishers just need better detectors. That's the wrong lesson, for two reasons.

First, detection tools answer the wrong question. They can estimate the probability that a passage was AI-generated. They cannot tell an editor whether that generation changed a finding, polished a sentence, or fabricated a citation. A detector flags style, not substance, and style is exactly the dimension that matters least for research integrity.

Second, and more fundamentally: you cannot detect your way out of an undefined category. If 70% of

journals have policies but disclosure sits at 0.1%, the gap isn't primarily an enforcement failure. COPE itself, in elevating undisclosed AI use to retraction-worthy misrepresentation, has effectively staked out a position that requires authors to know what counts as disclosable — and the underlying research hasn't caught up to tell anyone where that line sits. Asking detectors to enforce a boundary that publishers haven't agreed on is asking the wrong tool to do the wrong job.

What an Actual Response Looks Like

None of this is an argument for loosening disclosure standards, or for treating provenance ambiguity as somehow less serious than fraud because it's less malicious. An unverifiable research record is unverifiable regardless of intent. But the response has to match the actual shape of the problem, and right now it doesn't.

A few things would genuinely move this forward, and none of them are detection tools.

Define gradations, not a checkbox

"AI was used" is not a disclosure; it's an acknowledgment that disclosure is theoretically possible. Journals need disclosure categories that distinguish, at minimum, language polishing, literature synthesis, methods drafting, code generation, and data analysis because each carries a different verification burden, and lumping them together is exactly how the 9%-of-7,177 number happens. Authors aren't being asked a question they can answer honestly; they're being asked a question that doesn't map onto what they did.

Make provenance a structural field, not a sentence

A line in the acknowledgments is not provenance; it's a gesture toward provenance. The same logic that pushed scholarly publishing toward structured ORCID identifiers and mandatory data-availability statements over the last decade applies here; contribution-tracking needs to be a field in the submission system, not a sentence an author remembers to write or forgets to.

Stop treating the Integrity Hub model as the whole solution

Cross-publisher detection infrastructure is valuable and should keep expanding. It's genuinely effective against the fraud half of this problem. But its existence has let the field quietly substitute "we're working on AI integrity" for "we've defined what AI integrity requires." Those are not the same claim, and the PNAS data suggests the gap between them is where most of the actual exposure sits.

Accept that some of this will require new norms, not new rules

Rules without shared understanding of the category they govern produce exactly the outcome we're seeing — universal nominal compliance, near-zero actual disclosure. Disciplinary societies, not just publishers, need to be part of building the shared vocabulary for what AI contribution looks like in their specific kind of research, because a reasonable disclosure threshold in computational biology is not the same as one in clinical trials reporting.

The Crisis That Doesn’t Look Like a Crisis

Fraud is dramatic. It produces retraction notices, institutional investigations, and headlines. Provenance ambiguity produces none of that, which is precisely why it’s more dangerous in the way it accumulates. A field can point to its paper-mill takedowns and detection-tool donations and feel like it’s winning the integrity fight, while the foundational question — *can we actually trace how AI shaped the published record* — goes unanswered at scale, quietly, in most of cases.

The infrastructure built to fight fraud is necessary. It is not sufficient. And mistaking the second for an extension of the first is how a field ends up well-defended against the threat it already understood, and unprepared for the one it’s currently living inside.

Category

1. Thought Leadership
2. Trending Now

Date Created
2026/07/10
Author
anagha