Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 3

ISSN: 1540-3270

Generative Artificial Intelligence in Academic Research: Opportunities, Ethical Dilemmas, and Governance Pathways

[1]Dr. Rishi Kumar Sharma, [2]Dr. Mousam Singh, [3]Dr. Renu Bagoria, [4]Dr. Archana B Saxena

[1]Associate Professor, Jagannath University, Jaipur

[2]Associate Professor, Jagannath University, Bahadurgarh

[3]Professor, Jagannath University, Jaipur

[4]Professor, Jagan Institute of Management Studies

Abstract:

Generative artificial intelligence has moved into the academic world faster than academic norms, and so far the institutional reaction has been split into two moves: first, stating that such systems cannot be authors; and second, trying to detect the presence of such systems. This paper suggests that both responses are symptomatic responses and do not tackle the underlying structural problem. In scientific practice, authorship includes four elements which, historically, have gone together: intellectual contribution, accountability for correctness, credit, and provenance of warrant. Generative systems separate them, and the ban on machine authorship only addresses the accountability aspect, not the other parts of the problem, such as contribution, credit and provenance. We formalise this as the attribution decoupling problem and place it in the context of the warrant chain, a series of traceable human judgements that make a scientific claim answerable. We then argue that a system's governance should be calibrated for its use, on top of its use, and present two dimensions of a task-risk gradient: verifiability of the task's output by the researcher and closeness of the task to the scientific claim. The current uniform disclosure rules are seen as being both too broad for low risk uses and too narrow for high risk ones. We also claim that the use of detection-based enforcement is a dead end as it is impossible to have reliable detection equipment, and its errors are systematically distributed among researchers who write in a second language, making it an instrument of discrimination. Another approach is a governance based on process, declaration, provenance and verification, which takes place at four institutional levels. There are no invented statistics reported, seven propositions are provided and an analytical protocol.

Keywords: generative artificial intelligence; research integrity; authorship; peer review; research governance; scholarly publishing; provenance; academic ethics

1. Introduction

In just a few months, generative language models evolved from a novelty to a routine tool in various aspects of academic work. Researcher surveys suggest widespread use for drafting, editing, coding and literature work, while also showing a significant amount of concern regarding the integrity of the record with respect to such use of AI (Van Noorden & Perkel, 2023). Publishers, funders, and universities responded quickly, and that response was very similar, with the following elements: Generative systems can not be named as authors, since they have no responsibility for the work, and their use should be mentioned (COPE, 2023; ICMJE, 2024; Thorp, 2023). Several months after the technology was made public, the same view was taken in editorials in prominent journals (Nature, 2023; van Dis et al., 2023). These answers are right but not enough. The ban on machine authorship is a viable analysis: authorship involves accountability, and if a system cannot be held accountable, it cannot be an author. However, the ban does not address the practical issue, it simply addresses the definitional issue. The byline of a model does not answer the questions of the responsibility for a hallucinated citation it has generated, how to properly document the intellectual contribution it has made, whether it was possible to replicate the analysis it has performed, or how to evaluate the warrant of a claim that its derivation contained an opaque and indeterministic step. The challenge in the epistemics, the challenge in the credit ledger is addressed by the authorship policy.

A second dominant response - detection - is worse than being insufficient. Tools that claim to detect machine-generated text have not performed well under realistic conditions, as the error rates make single decisions unsafe (Weber-Wulff et al., 2023), and theoretical research indicates that there may be no reliable way to do so generally as generation improves (Sadasivan et al., 2023). More seriously, errors in the detectors are not uniformly distributed. For non-native English writers, the rates for which they are detected as machine-generated tend to be moderately high, as the features used by the detectors (such as limited lexical variation and conventional syntax) are common to second-language writing and model output (Liang et al., 2023). An enforcement system based on these would be especially burdensome for researchers in just those places where the technology's equity advantages are greatest.

This paper develops an alternative. It has three steps to its argument. First, it points to structural issue called attribution decoupling, which is a separation of the concepts of contribution, accountability, credit and provenance that authorship conventions once combined. Second, it maintains that risk is not an inherent feature of generative systems, but a feature in the context of the research process, and it proposes a two-dimensional task-risk gradient for revealing risk, one dimension referring to task complexity, and the other to the risk disclosure process. Thirdly, it identifies governance routes at four institutional levels and shows that process-based instruments focused on declaration, provenance and verification are the only ones that will stand up to a technology that is not easily detectable.

The paper also rejects two terms which are often used in public discourse. It doesn't see generative AI as primarily a cheating issue, as the most dangerous scenarios are when researchers are over-trusting the output of a plausible AI model, rather than when they're doing deliberate wrongdoing. It doesn't imagine the technology is unqualified for the task of discovery, as the means by which it accelerates output differ from the means by which it produces knowledge, and the gap between means and ends is itself a governance issue (Messeri & Crockett, 2024).

2. Capabilities and the Research Workflow

2.1 What these systems are, and why it matters for governance

Governance based on a flawed concept of the nature of technology will fail, so a short technical explanation is in order. Modern generative language models are trained to generate the next token in a very large text corpus, and are then fine-tuned to follow instructions and to produce samples that are found most useful by human evaluators (Bommasani et al., 2021; Brown et al., 2020; Ouyang et al., 2022). There are three consequences, all with a governance implication.

First of all, fluency and accuracy are two distinct properties. The training goal is to get a plausible, not correct, continuation for the word, and a plausible false continuation is not a bug to be patched, it is a characteristic of the generation task as an optimised system (Ji et al., 2023). The governance implication is that outputs do not need to be checked for signs of unreliability as the usual "rules of thumb" for detecting unreliable text, such as hedging, awkwardness and vagueness, are precisely what is "tuned" out. Some previous studies have revealed that domain experts often struggle to tell generated scientific abstracts apart from original ones (Else, 2023; Gao et al., 2023). Secondly, instruction tuning is an optimization that works for a rater's preference rather than the truth. Systems, therefore, are more likely to give agreeable, confident, conventional answers and this is important when using the system in research since the skill of a good reader is disagreement – a system tuned to be helpful is a poor adversarial check on a researcher's own reasoning. Third, there are non-deterministic outputs and there are models that are versioned, updated and ultimately removed. Because the same prompt to the same nominal system can result in different output materially at different times and a model used in a study may not be available in a similar form when it is replicated, it is found that material differences can arise when the two are compared.When comparing material differences arise because the same prompt to the same nominal system can result in different output materially at different times and a model used in a study may not be available in a similar form when it is replicated. It is not merely a matter of the inconvenience of being able to reproduce the results of a model-assisted step; it is a matter of the fact that without explicit provenance recording the results of a model-assisted step are permanently unauditable where the results of a documented statistical procedure are not.

2.2 Mapping capability to workflow

Any useful analysis must begin by disaggregating what these systems do, because treating them as a single instrument produces policy that is wrong for most uses. Table 1 maps generative capabilities onto stages of the research workflow and identifies the failure mode characteristic of each.

Table 1. Generative capabilities mapped to the research workflow

Research stage	Typical use	Genuine opportunity	Characteristic failure mode	Detectability of failure by the researcher
Problem formulation	Ideation, framing, gap identification	Broadens the option set beyond the researcher's priors	Regression to conventional framings present in training data	Low; the absence of an idea is not observable
Literature work	Search, screening, synthesis	Substantial time saving in screening at scale	Fabricated or misattributed references; confident misstatement of findings	Moderate; requires checking each source
Study design	Protocol drafting, power considerations	Surfaces standard design elements a novice would omit	Plausible but inappropriate design transfer across contexts	Moderate; depends on methodological training
Data collection	Instrument drafting, coding frames	Faster iteration on instruments	Subtle construct drift in item wording	Moderate
Analysis	Code generation, statistical scripting	Large productivity gain where output is testable	Silently incorrect implementation that runs without error	High, if and only if the code is tested
Interpretation	Explaining results, framing implications	Articulates alternatives the author had not considered	Confident over-interpretation; illusion of understanding	Low

Research stage	Typical use	Genuine opportunity	Characteristic failure mode	Detectability of failure by the researcher
Writing	Drafting, editing, translation	Major equity gain for second-language writers	Homogenisation of expression; smoothing over uncertainty	High for language, low for hedging loss
Peer review	Drafting reports, summarising submissions	Reduces reviewer burden	Confidentiality breach; superficial or fabricated critique	Low to the editor; invisible to the author

Two observations follow from the table. The first is that opportunity and risk are not inversely related across the workflow; they are highest together in literature work and analysis, which is why blanket prohibition and blanket permission are both poor policies. The second is that detectability of failure by the researcher varies enormously, and it is this variation, rather than the intensity of use, that should drive governance. A researcher who generates code and then tests it has a verification mechanism; a researcher who accepts a synthesis of literature they have not read does not.

2.3 Opportunities taken seriously

A framework structured around risk should make the positive case with the same level of clarity, otherwise a governance mechanism will be overly constraining when it fails to take into account the benefits that are forgone. People in four opportunities are not just uncertain, but significant.

The first is the decrease of the language penalty. The researchers in a second language have long been a disadvantaged group in a system that is largely dominated by English.The systematic disadvantage plaguing researchers in a second language has been well known for many years in a publication system conducted almost exclusively in English. Language assistance directly tackles it at low cost and at the place of its binding. The technology provides the greatest benefit to academic research, even though it is disproportionately benefitting those in lower income and non-Anglophone welfare systems, and it is being put at risk by the enforcement on detection, for the reasons outlined in Section 6.

The latter is scaling up screening. Identifying which of several thousand records are worth reading in full is a mechanical process that has a definitive result – one can audit a sample of those to be included and excluded and estimate the error rate and compare that with their own judgement. This is the paradigm case for a task that is both highly verifiable and moderately close to the claim, and to equate it with the production of the synthesis is the most obvious case of over-inclusiveness arising from uniform policy (Wagner et al., 2022).

The third one is the reduction of methodological barriers. A researcher with no programming background can create analytical code, while one without a statistical background can get an explanation for an unknown procedure. This really does open up avenues to methods, and it also introduces the danger that the ability to produce an analysis outstrips the ability to assess it: the verifiability dimension in Section 4 is a bit of a property of the researcher, rather than a property of the task. The correct answer is not to be able to withdraw the capability, but to be able to check it at the same time.

The fourth type is adversarial use – the least talked about and probably the most valuable. A system that is required to produce the strongest objections to an argument, alternative explanation of the result, or reasons why a design

might fail, does a job that few and overworked colleagues do well and usually do not do at all. This use is low-risk in both respects: the output is not a claim; nothing goes into the manuscript without being examined. An undifferentiated spirit of caution would deprive it of the application most closely associated with an organised scepticism and allow other more dangerous applications.

3. The Attribution Decoupling Problem

3.1 What authorship bundles

The system of scientific authorship is an uncommon one which serves a number of functions in a single stroke. Table 2 separates them. Who did the intellectual work is recorded in contribution records. Accountability creates responsibility as to who is responsible for the incorrect work. Credit gives rewards, which provides base to career, and funding. Provenance provides the link of warrant by which one can trace a claim back to its judgement. It has long been recognised that these bundles cause difficulties without any technology, so proposals for replacing authorship with contributorship and the development of structured contribution taxonomies (Allen et al., 2014; Brand et al., 2015; Rennie et al., 1997) were spurred on.

Table 2. The four components bundled in authorship and the effect of generative AI

Component	Function	Pre-AI carrier	Effect of generative AI	Adequacy of the current policy response
Contribution	Records who performed the intellectual work	Author list; contribution statements	A non-author performs work previously done by an author	Inadequate; contribution statements have no category for it
Accountability	Establishes who answers for error	Authorship implies answerability	Cannot attach to the system; must remain with the human	Adequate; this is what the authorship prohibition secures
Credit	Allocates reward and reputation	Byline order; citation	Credit accrues to the human for work they did not do	Not addressed at all
Provenance	Traces the claim to the judgement producing it	Methods section; data and code availability	Model version, prompt, and non-determinism are typically unrecorded	Largely unaddressed; no standard reporting exists

The ban on machine authorship guarantees the second but not second part. This is NOT an indictment of the prohibition; it is just a note that the policy discussion ended with resolving the simplest of the four issues. There is no vocabulary available in contribution statements for machine involvement. More and more credit for output is being given to people who did not create it, leading to an increasing perversion of the reward mechanism which will continue to grow. Although it is the most important for trustworthy literature, there has been no research into

provenance, the topic has received the least attention, and it is the one for which there are already well-developed tools for the other fields that have adopted provenance, such as model documentation in machine learning (Gebru et al., 2021; Mitchell et al., 2019).

3.2 The warrant chain

Figure 1 presents the structure that the decoupling threatens. A scientific claim is warranted not because it is true in isolation but because an unbroken chain connects it to evidence through judgements for which identifiable people are answerable. Merton's account of organised scepticism describes the social mechanism that maintains this chain: claims are held provisional until independently scrutinised, and scrutiny requires that the steps be inspectable (Merton, 1973).

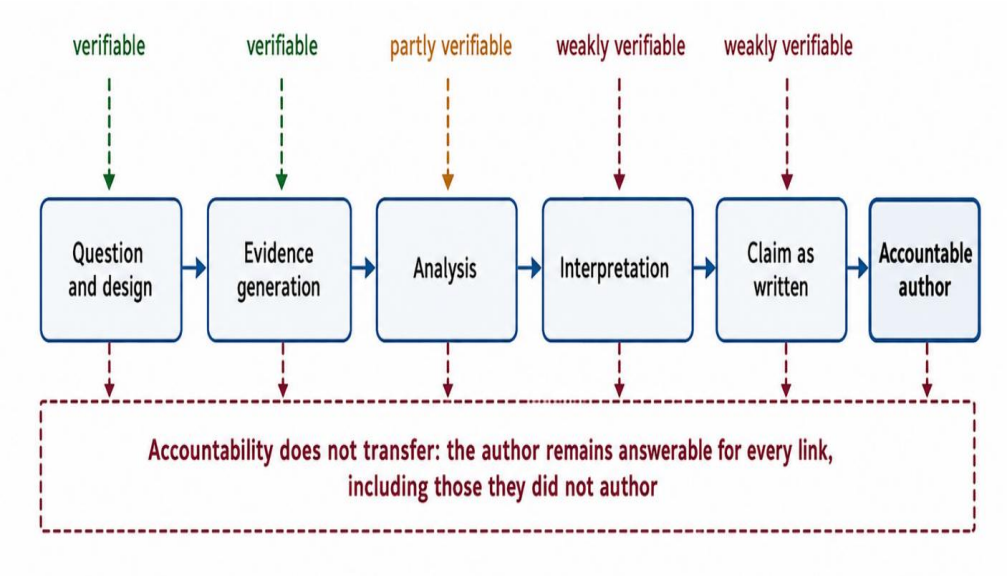

Figure 1. The warrant chain and generative model insertion points. Risk rises along the chain: insertions closer to the scientific claim are less verifiable and enter the claim more directly, while accountability remains with the human author at every link.

Generative systems can be inserted at any link. What matters is that the insertion does not transfer accountability, which remains with the author regardless, while it may remove verifiability, which the author needs in order to discharge that accountability. This produces a specific and novel form of research risk: authors become answerable for claims whose derivation they cannot fully inspect, not through negligence but because the instrument does not expose its reasoning. The problem is structurally similar to reliance on a proprietary instrument or a black-box statistical package, with the important difference that generative output is fluent, plausible, and therefore actively resistant to the scepticism that unfamiliar output would attract.

4. The Task-Risk Gradient

4.1 Two dimensions

If risk is based on position in the warrant chain, then governance should also be based on the position in the warrant chain. We suggest two measures. The first is verifiability – the ease and accuracy with which the researcher can determine if the results are accurate. The generated code is extremely verifiable, as it can be tested with known examples; a synthesis of 40 papers read by the researcher isn't verifiable without the researcher doing the work the synthesis was supposed to do. The second one is claim proximity - the degree to which the output enters the scientific claim. The claim is the interpretation of results, reference formatting is remote. Tasks representative of this plane are identified in Figure 2, and the disclosure regime that each quadrant suggests in Table 3.

Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 3

ISSN: 1540-3270

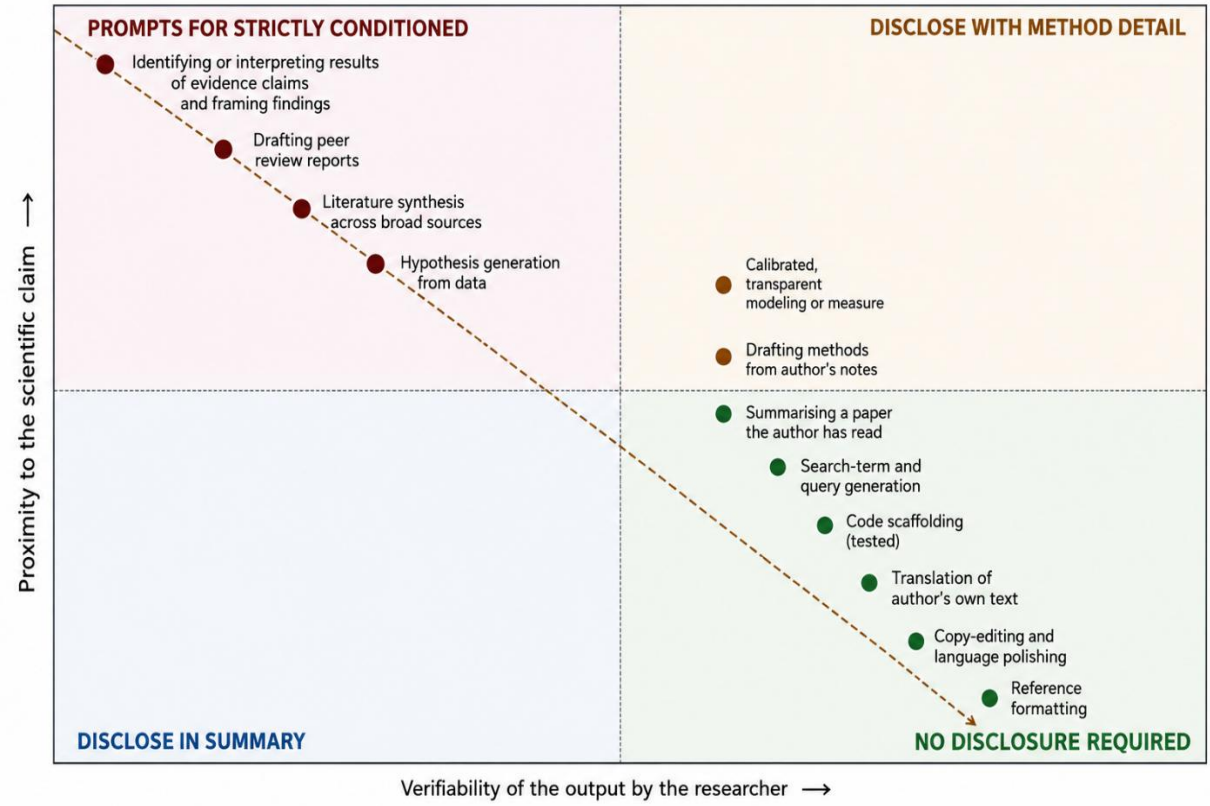

Figure 2. The task-risk gradient. Research tasks positioned by verifiability of output and proximity to the scientific claim, with the implied calibrated disclosure regime by quadrant. Positions are analytic judgements offered for discussion, not measurements.

Table 3. Calibrated disclosure regimes by quadrant

Quadrant	Verifiability	Claim proximity	Representative tasks	Disclosure regime	Rationale
Q1 Benign	High	Low	Copy-editing, formatting, translation of the author's own text	None required	Functionally equivalent to existing tools already exempt from disclosure
Q2 Instrumental	High	High	Code generation, scripted analysis	Disclose with method-level detail: model, version, and verification performed	Verifiable, but enters the claim; reproducibility requires the record
Q3 Assistive	Low	Low	Ideation, search-term generation, first-pass screening	Summary disclosure	Errors do not propagate into the claim, but

Quadrant	Verifiability	Claim proximity	Representative tasks	Disclosure regime	Rationale
					shape the option set
Q4 Hazardous	Low	High	Literature synthesis of unread sources, evidence claims, interpretation, peer review	Prohibited, or permitted only with independent human verification of every element	Author cannot discharge accountability for what they cannot verify

The following three aspects of this scheme should be highlighted. It first highlights the reasons why the existing uniform policies are not working for anyone. Always requiring disclosure is over-inclusive in Q1 where disclosure statements contain no information, and has no signal due to its ubiquity. It is also under-inclusive in Q4, where what is needed is not disclosure, but verification, and a reader who is told that it was a literature synthesis based on a model is not better equipped to believe in it. Secondly, the gradient is not broken and the quadrants are not a classification but rather an exposition tool. Third, the location of particular tasks is debatable and ought to be debated; the claim of the framework is on the measurement, not on the location of an individual task.

A corollary, and one that is more practical, is the asymmetry of the second dimension. In part, verifiability is a quality of the task, and in part a quality of the researcher. An experienced methodologist can check a generated analysis plan not an experienced doctoral student. This means that the same use, recorded by different researchers at different locations can have different risk, and is highly uncomfortable when dealing with rule-based governance, but is exactly the type of judgement that supervision and training is meant to provide.

4.2 Discipline-specific calibration

Both dimensions are general, and the position of a specific duty within them is not, and a governance system that is oblivious of discipline will be out of kilter in predictable ways. It's illustrated with three contrasts.

There is a lot of output from computational and quantitative models that can be checked by running the model. If generated code fails to produce the expected result for known cases, the result is not considered to be verifiable, thus shifting a significant portion of the workflow towards the right on the verifiability axis, which makes permissive policy with method-level reporting appropriate. This is the reason that in Q2 the disclosure requirement should be to describe the verification, not the use of a system, as there is a risk that a system may be used without causing a failure.

In subjects where the emphasis is on literature and theory, it is the other way round. The method and the claim of scholarship synthesis are located on the edges of the verifiability and claim proximity spectrum. Research on the application of automated technique in literature reviews has focused on the importance of the judgment about relevance, quality and conceptual relation, which none of the automated technique can check without redoing the literature review (Wagner et al., 2022). It has to be governed in a similarly strong way and disclosure is not enough for the reasons above.

Qualitative and clinical research is an exception, with the binding constraint not being epistemic primarily. The transcripts of the interviews, clinical notes and field materials include identifying information about those who agreed to a specific use of the data, and it is a matter of data-protection and research ethics before it's a matter of integrity. Ethics approvals that were obtained prior to the existence of these systems do not envision such transfer, and researchers often aren't aware that a processing step has taken place. This is one of the most obvious instances where institutional ethics committees, and not the journal, has the real leverage: it is the appropriate governance instrument that is not disclosure at publication, but rather, a condition at approval.

5. Ethical Dilemmas Beyond Authorship

Table 4 sets out six dilemmas that the authorship debate has crowded out. Each is briefly elaborated below.

Table 4. Ethical dilemmas and their governance implications

Dilemma	Nature of the problem	Who bears the harm	Why current policy is insufficient	Governance implication
Fabricated references and evidence	Fluent invention of plausible sources and findings	Readers; the cited who are misquoted; the record	Disclosure does not correct the error	Desk-stage reference verification; author attestation of source access
Confidentiality in peer review	Submitting unpublished manuscripts to third-party systems	Authors, whose unpublished work is disclosed	Most policies address authorship, not reviewing	Explicit reviewer prohibition; secure institutional deployment
Reproducibility and provenance	Non-determinism, undisclosed versions, model deprecation	Anyone attempting replication	No reporting standard exists for model-assisted steps	Structured reporting of model, version, prompt, and date
Equity	Access to capable systems; language advantage and detector bias	Researchers in under-resourced settings and second-language writers	Detection-based enforcement inverts the equity gain	Prohibit detector-based sanction; support equitable access
Volume and paper mills	Cost of producing plausible manuscripts falls sharply	Editors, reviewers, and the integrity of the record	Screening capacity does not scale with submission volume	Provenance requirements; submission-level integrity checks
Epistemic dependence	Fluent output suppresses scepticism; understanding is simulated	The discipline, through unnoticed narrowing	No policy instrument addresses it	Training; deliberate verification practices; methodological education

This is the best documented failure: fabricated references. Research on the bibliographies produced by language models has revealed high rates of falsified and/or inaccurate citations, even plausible, but non-existent author-title-journal combinations (Bhattacharyya et al., 2023; Walters & Wilder, 2023). The above described phenomenon has been well described in the technical literature as hallucination, which is not a bug to be fixed but a property of the generation objective (Ji et al., 2023). The level of disclosure is an insufficient response since the harm is likely to happen with or without disclosure; what is necessary is to be assured that every source cited is present and states exactly what is claimed of it.

The issue of confidentiality of peer review has not been addressed as much as it should be. A reviewer for a commercial system that submits a manuscript to the system for a report has revealed unpublished work to a third party, possibly for further training. There have been analyses of review texts that have detected proof of significant model assistance in certain contexts (Liang et al., 2024); and discussions in the research-integrity literature that support the need to differentiate between authorship rules and apply explicit and separate principles to reviewing (Hosseini & Horbach, 2023).

A special type of lack of reproducibility. Generative outputs vary across versions and models are updated and retired and the same prompt can result in significantly different output across versions. Reproducibility is a characteristic of computational research that assumes that a given procedure will yield the same result (Munafò et al., 2017; Nosek et al., 2015), and in the field of machine learning, reporting practices have also struggled with this problem (Gundersen & Kjensmo, 2018; Pineau et al., 2021). A specific and tractable gap is the lack of a similar standard for model-assisted steps in disciplines other than computing.

There are two-sided cuts and a two-sidedness. The reduction of a longstanding disadvantage of researchers working with a second language is one of the most obvious welfare benefits of language assistance. On the other hand, use of the most capable systems is unevenly distributed and detector-based enforcement provides punishment to the same population it provides benefit to. (Liang et al., 2023) Such a governance system would then turn an equalising technology into an amplifying system for the existing disadvantage.

Structural challenge with volume. The cost of generating a believable manuscript is now much lower than ever before, the capacity of the systems of editorial and review has stayed unchanged, and organised paper-mill activity has existed long before the technology has and is well placed to use it (Byrne & Christopher, 2020; Else & Van Noorden, 2021). The pertinent defense is not that the text is detected but the things that a mill cannot easily fabricate: institutional affiliation, data provenance, and ethical approval.

Epistemic dependence is the most intractable, and perhaps the most-important. Analysis of recent literature suggests that the tools used in AI can create the illusion of understanding, as the breadth of what is known seems to grow, but the breadth of approach and of question shrinks, leading to monocultures of method and question (Messeri & Crockett, 2024). Previous studies have warned of the impact of LLM on the epistemic practices of science in general, on the basis of its shortcomings as a language generation system (Bender et al., 2021). There is no disclosures requirement that deals with it, since there is nothing to be hidden.

6. Why Detection-Based Governance Fails

The intuitive response to an integrity threat is enforcement, and the intuitive instrument is detection. Table 5 sets out why this fails, distinguishing product-based governance, which examines the output for evidence of machine origin, from process-based governance, which requires that the process be declared, recorded, and verifiable.

Table 5. Product-based versus process-based governance

Dimension	Product-based (detection)	Process-based (declaration and provenance)	Assessment
Technical reliability	Poor and declining as generation improves	Independent of generation quality	Decisive advantage to process
Error distribution	Systematically biased against second-language writers	No comparable bias	Decisive advantage to process

Dimension	Product-based (detection)	Process-based (declaration and provenance)	Assessment
Incentive created	Evade detection; avoid disclosure	Record accurately; verify before submission	Process aligns incentive with integrity
Coverage	Text only; silent on analysis, synthesis, and review	Covers any stage where a declaration is required	Process covers the higher-risk stages
Cost of enforcement	Continuous arms race with generation	One-time infrastructure plus routine compliance	Process is cheaper over any horizon
Failure mode	False accusation of honest researchers	Undeclared use by dishonest researchers	Both real; the second is less unjust
Compatibility with watermarking	Would improve if watermarking were universal	Complementary; provenance is strengthened by it	Watermarking helps both, but requires provider cooperation

The last row is worth a note. Technically, it is possible to cryptographically watermark generated text, and this has been shown to be possible (Kirchenbauer et al., 2023), and if this is done universally by providers, this would improve the provenance of generated text materially. It's not, however, a replacement for process governance as it is voluntary, can be paraphrased, doesn't mention other uses than the textual one, and isn't present in open-weight systems that any researcher can run locally.

The second argument against detection is normative rather than technical. An integrity regime that generates fake news, that is mostly written by researchers who are not native speakers of the language, is not only ineffective, it has serious consequences for the persons involved for a characteristic that has nothing to do with standards of integrity. This would not be cured by even significant advances in detector sensitivity, since the probability of misconduct is small, and the population toward which the misconduct could go is known and is already disadvantaged. This is a good enough reason to deny the detector output as a basis of sanction without any improvement in the technical quality.

Finally, institutional cost is to be discussed. The nature of detection regimes means that they impose an ongoing cost, which increases as they are used more, and as detection becomes more sophisticated, and they create an adjudication burden because flagged manuscripts must be evaluated by humans, author responses prepared, and often, appeals filed. Process regimes involve an up-front infrastructure cost, a small ongoing compliance cost, and disputes over known facts, not probabilistic interpretations of the text. Therefore, institutions currently investing in detection are in effect buying a ‘liability' instead of making a ‘capital improvement' and at a time when the technical ground for the purchase is still being challenged (Sadasivan et al., 2023; Weber-Wulff et al., 2023).

7. Governance Pathways

Figure 3 organises available instruments across four institutional levels and three instrument classes. Table 6 sets out the leverage and characteristic failure of each level.

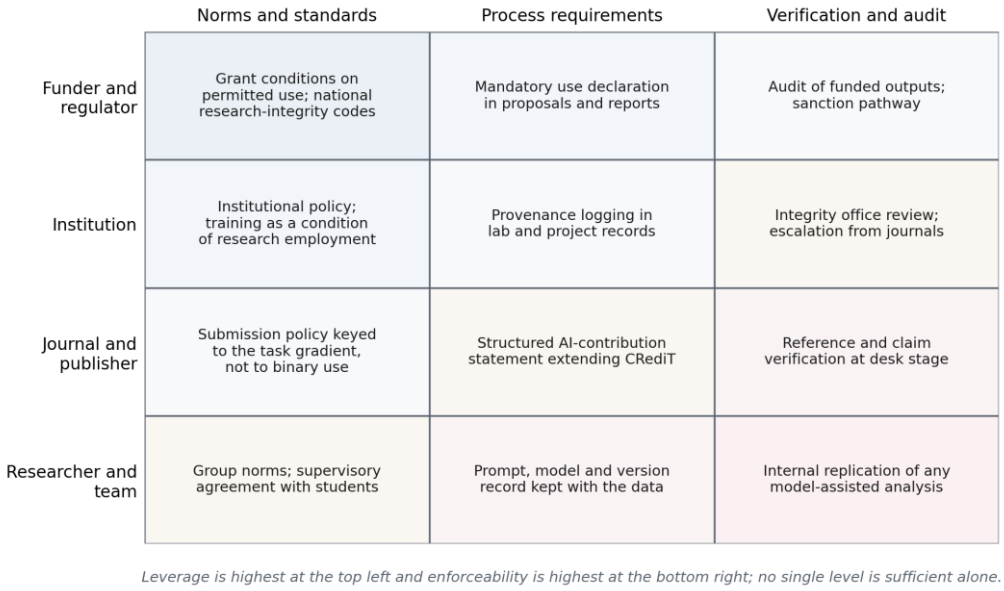

Figure 3. Governance instruments by institutional level and instrument class. Leverage is highest at the upper levels and enforceability highest at the lower ones; no level is sufficient alone.

Table 6. Governance levels, leverage, and failure modes

Level	Principal instrument	Leverage	Characteristic failure	Precondition for effectiveness
Researcher and team	Group norms; provenance records kept with data	Highest proximity to the actual conduct	Unenforceable; varies with supervisor practice	Training and a genuine expectation of scrutiny
Journal and publisher	Submission policy; structured contribution statements	Controls access to the reward system	Policy proliferation; inconsistent across venues	Convergence on a common declaration schema
Institution	Employment policy; integrity office; training requirement	Can impose consequences; holds the employment relationship	Slow; risk-averse; reactive to external complaint	Resourced integrity function with investigative capacity
Funder and regulator	Grant conditions; national codes; statutory instruments	Sets the terms all other levels operate under	Blunt; slow to revise; may over-restrict beneficial use	Calibration to the task gradient rather than to use per se

There are 3 design rules. The first is calibration. It is not good governance to require a policy on the use of the system but rather on the kinds of tasks performed on the system and how the output was checked. The second is that this is the load-bearing instrument, which is called the “provenance.” A documented description of model, version, date, prompt or prompt class, and verification conducted, kept with the project data, covers most of the needs for disclosure in the model and is compatible with existing open-science infrastructure (Nosek et al., 2015;

Wilkinson et al., 2016). The third is that contribution taxonomies need to be expanded and not eliminated. Expanding the CRediT framework to include this machine assistance modifier on the corresponding roles will be a smaller and more manageable change than the creation of a parallel disclosure system as suggested by Allen et al. (2014) and Brand et al. (2015).

International instruments offer a framework and not specifics. The UNESCO recommendation on AI ethics lays down principles that are also applicable to research contexts (UNESCO, 2021), while the EU's regulation on AI includes specific transparency obligations which are applicable to generated content, but with specific carve-outs applicable to research use, which should be checked against the current text (European Union, 2024). Neither provides operational guidance like a methods section, which is the purview of the disciplinary and editorial bodies.

7.1 A minimum viable declaration schema

To be calibrated, there must be something to take exception to, and a lack of common schema is a major reason that the current policies make uninformative statements. A workable minimum has five fields, and should only take less than a minute to complete for most submissions. The first one is the task class which is from a short controlled vocabulary, not free text as free text statements are not analysable at scale and are not the purpose of collecting the statements. This is the system and system version; the date of use is not included, otherwise no attempt to replicate could be made. The third is the checking which was done: sources were checked against originals, code was run on test cases, analysis reproduced independently, or not done. The fourth relates to who did that verification, which reintroduces that responsibility link – something the current settlement does make clear about – and should be made explicit. The fifth only comes into force if confidential or personal material was processed, and records the documents submitted and under which approval.

There are two intentional characteristics of this schema. It does not inquire of how much of the text has been generated, and can't be answered, is not a matter of risk, and is the substantive disclosure, not the use field. A submission that shows substantial use, but is backed up with documented evidence is more credible than one that shows minimal use, but has no evidence. To test several of the propositions in Section 8 (Allen et al., 2014), the data produced by this task-class vocabulary needs to be analysable across the literature, which requires an alignment of the task-class vocabulary with existing contribution taxonomies.

8. Propositions and Analytical Protocol

Table 7 states seven propositions with the evidence that would test them. The framework is intended to be falsifiable, and several of these are testable with data that publishers and integrity offices already hold.

Table 7. Propositions and testing strategy

Proposition	Statement	Testing strategy	Parameter of interest
P1	Disclosure rates fall as policies become more uniformly demanding, because ubiquitous low-risk disclosure devalues the act of disclosing.	Compare declaration rates across venues with uniform versus calibrated policies	Policy-type contrast [b1]
P2	Error incidence is a function of the verifiability of the task, not of the intensity of model use.	Audit of published errors against reported task type	Coefficients [b2v], [b2i]
P3	Detector-based screening produces false-positive rates that vary	Audit of screening outcomes against authorship metadata	Differential rate [b3]

Proposition	Statement	Testing strategy	Parameter of interest
	systematically with author language background.		
P4	Fabricated-reference incidence falls sharply under desk-stage verification but not under disclosure requirements alone.	Natural experiment across journals adopting one, both, or neither	Treatment contrast [b4]
P5	Structured provenance records improve replication success for model-assisted analyses relative to narrative disclosure.	Replication attempts on matched samples	Replication rate difference [b5]
P6	Model assistance in writing reduces the language-based disadvantage of second-language authors in editorial outcomes.	Desk-rejection and acceptance rates before and after availability, by author language	Difference-in-differences [b6]
P7	Concentration of method and topic increases with model use at the field level, consistent with monoculture concerns.	Diversity indices on published methods and questions over time	Trend in diversity [b7]

Four measurement cautions apply. Self-reported use is subject to strong social desirability pressure while norms remain unsettled, so declaration rates should not be read as use rates. Author language background is not directly observable and its proxies are imperfect and themselves sensitive, requiring careful ethical review before any such audit. Published error incidence is censored by what reviewers catch, so audits should sample verification independently rather than rely on published corrections. And any comparison across venues confounds policy with the disciplinary and quality composition of submissions, so within-venue policy changes provide considerably better identification than cross-sectional comparison.

9. Discussion

9.1 Implications

The key take out from the framework is that the governance discussion has been focused on the wrong thing. When policies are written as such, they will be unenforceable, not well-understood, and not equitable. The verifiability of the warrant chain as the regulated object leads to policies that can be enforced when they are submitted, that inform the reader, and that are neutral with respect to who submitted them. This reframing also accounts for the fact that the biggest trouble with the technology isn't misconduct trouble. A researcher who takes a fluent and incorrect synthesis is not a cheat, but rather a victim of training and verification failure to an instrument whose failure mode is not visible.

A second implication is to do with opportunities, rather than under-estimating them, I have made it my point not to overestimate them, either, in this paper. The outcomes of screening large literatures and drafting a code, and of reducing the language penalty, are real and significant gains, and the equity gain is the best positive argument for permissive policy. However, the process by which these benefits come is a decline in the costs of providing scientific text and analysis, and the speed of knowledge production is not equal to the speed of text production.

The first is more important than the second where they differ, so that the Goliath of governance safeguards the David. This is the essence of what P7 is about and an efficiency-only evaluation of the technology will fail to capture.

9.2 Objections

There are three objections that need to be directly addressed. The first is that the framework over-engineers a problem which the ordinary rules of academic judgment - which have historically guided research with tools - can solve: researchers always used tools, always have been accountable for the outcomes of their research using tools, and no new apparatus is needed. This objection has some force in regard to Q1 and much in regard to Q2; and the framework gives this to them in regard to Q2 by imposing no requirement. It fails for Q4 since that is not a novelty, but rather a difference between the previous tools and this tool is that it is fluently used and also unverifiable. While the output of a calculator can be checked and a search engine returns sources that exist, a synthesis of unread literature is neither and has the unintended consequence of generating confident, well-formed, entirely fictional references at scale (Walters & Wilder, 2023).

The second is that calibration is impractical, for researchers would just classify their use in the quadrant that has the lightest obligation. This is a real drawback, and that's why in Section 7.1, the schema is not one of task classification, but one of verification. So a task that is misclassified with an accurate verification record provides the information a reader will need, but a well classified task with no verification record does not. But even more strongly for the alternative, because binary disclosure regimes are gameable just by failing to disclose (and in an absence of schema to be inconsistent with).

Third, the prohibition is impossible to enforce in Q4, and will push the use of it further underground, compounding the informational situation. This is right and that is why the framework doesn't just rely on prohibition. It is not so much a rule as a rule of thumb: No part of the output should be placed in the manuscript without being verified by a named person, this can be enforced at the desk stage in relation to references and at the review stage in relation to claims, and it is not dependent on the fact that a system was used or not. The researcher who validates all the requirements of the process, regardless of what they are using, the correct result, which the detection-based regime can't achieve.

9.3 Limitations

Three restrictions need to be clearly and simply said. It is a conceptual/ normative document, whose propositions are not measured, but rather drawn; the positions assumed to tasks in figure 2 are analytic judgements that may be challenged, not measurements of anything. Second, the field is changing quickly, and the technical statements of the detection and policy positions of publishers and regulators should be checked against current sources before relying on them; several of the instruments mentioned were changed in the time frame of this analysis. In addition, the analysis is heavily based on the Anglophone journal system and its integrity institutions, and it is possible that much less of it could transfer to research systems that have different publication incentives, less developed integrity institutions, or a stronger incentive to focus on volume, where the paper-mill and volume risks discussed in Section 5 may be tighter.

10. Conclusion

Generative AI is not an ethically new paradigm for academic research. It gives it one that is familiar, the meaning of warrant, where it is difficult to maintain the traceability. The institutional response thus far has been to establish the fact that a system is not an author, the easiest part of the problem; and to pour much of the effort into the least promising instrument, by trying to detect something that will not likely be detected, and inflicting the error on the least likely to bear it.

The alternative that has been developed here is a question that a researcher is asked: Was there any evidence involved, but every link between evidence and claim is one that could be answered by a person, and could be checked by another person? That question can be answered, it is factored by risk, it is not a technical arms race, it is not a question of style of prose. It also calls on institutions which are not detected, such as building

infrastructure for provenance and training in routine verification. This is not something that can be accomplished by simply checking a box in a submission portal, which is why it hasn't been done yet. The propositions, and the protocol offered here, are designed to make the case testable, rather than simply asserted, so that it is refutable.

References

[1] Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature, 508*(7496), 312-313.

[2] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

[3] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). *On the opportunities and risks of foundation models* (arXiv:2108.07258). arXiv.

[4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

[5] Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus, 15*(5), e39238.

[6] Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics, 5*(5), 277-280.

[7] Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing, 28*(2), 151-155.

[8] Byrne, J. A., & Christopher, J. (2020). Digital magic, or the dark arts of the 21st century: How can journals and peer reviewers detect manuscripts and publications from paper mills? *FEBS Letters, 594*(4), 583-589.

[9] COPE. (2023). *Authorship and AI tools: COPE position statement*. Committee on Publication Ethics.

[10] Else, H., & Van Noorden, R. (2021). The fight against fake-paper factories that churn out sham science. *Nature, 591*(7851), 516-519.

[11] Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature, 613*(7944), 423.

[12] European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

[13] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92.

[14] Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *npj Digital Medicine, 6*, 75.

[15] Gundersen, O. E., & Kjensmo, S. (2018). State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1).

[16] Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review, 8*, 4.

[17] Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2023). Using AI to write scholarly publications. *Accountability in Research*. Advance online publication.

[18] ICMJE. (2024). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals*. International Committee of Medical Journal Editors.

[19] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1-38.

[20] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*.

[21] Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., ... Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*.

[22] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779.

[23] Merton, R. K. (1973). The normative structure of science. In N. W. Storer (Ed.), *The sociology of science: Theoretical and empirical investigations* (pp. 267-278). University of Chicago Press.

[24] Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627*(8002), 49-58.

[25] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229).

[26] Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021.

[27] Nature. (2023). Tools such as ChatGPT threaten transparent science: Here are our ground rules for their use [Editorial]. *Nature, 613*(7945), 612.

[28] Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425.

[29] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730-27744.

[30] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research. *Journal of Machine Learning Research, 22*(164), 1-20.

[31] Rennie, D., Yank, V., & Emanuel, L. (1997). When authorship fails: A proposal to make contributors accountable. *JAMA, 278*(7), 579-585.

[32] Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-generated text be reliably detected?* (arXiv:2303.11156). arXiv.

[33] Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science, 379*(6630), 313.

[34] UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific and Cultural Organization.

[35] van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature, 614*(7947), 224-226.

[36] Van Noorden, R., & Perkel, J. M. (2023). AI and science: What 1,600 researchers think. *Nature, 621*(7980), 672-675.

[37] Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology, 37*(2), 209-226.

[38] Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*, 14045.

[39] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*, 26.

[40] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018.