OPEN ACCESS

Artificial Intelligence COMMENTARY

From vulnerable AI referees to accountable AI-assisted peer review

Yu Wu[1,2,*]

[1]National Academy of Forestry and Grassland Administration, Beijing 102600, China

[2]Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081, China

*Correspondence: wuyu20@mails.ucas.ac.cn

Received: June 14, 2026; Accepted: October 14, 2026; Published Online: October 15, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100068

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Citation: Wu Y. (2026). From vulnerable AI referees to accountable AI-assisted peer review. *The Innovation Informatics* **2**:100068.

Wang et al. provide a timely security-centered assessment of artificial intelligence (AI)-assisted peer review.[1] By examining authority bias, assertion strength, rebuttal sycophancy, and contextual poisoning, the study shows that AI referees are not merely imperfect text generators, but evaluative systems whose judgments can be systematically shifted by signals that should be irrelevant to scientific merit. Their experiments show, for example, that prestigious affiliations can raise scores, cautious wording can be penalized, confident but evidence-free rebuttals can shift evaluations upward, and manipulated contextual information can alter judgments.[1] These vulnerabilities are consistent with independent evidence: invisible prompt injections embedded in manuscripts have been shown to inflate AI review scores and raise acceptance recommendations from 0% to nearly 100% in simulated medical peer review, while controlled studies have also documented affiliation bias favoring authors from highly ranked institutions.[2,3] Together, these findings move the debate beyond generic concerns about hallucination toward a more precise question: under what procedural conditions can AI-assisted reviewing be trusted?

Reliability should therefore include procedural robustness, which we define as the capacity of an AI-assisted review process to preserve substantively consistent and evidence-based judgments when exposed to normatively irrelevant or adversarial changes in author identity, rhetorical framing, rebuttal style, prompts, or contextual information. Procedural robustness has at least four practical dimensions: resistance to adversarial or irrelevant cues; consistency across repeated or reformulated evaluations; fairness across authors and institutions; and traceability from evaluative claims back to supporting evidence. This definition distinguishes robustness from mere linguistic fluency or apparent expertise. A system may write an articulate review while remaining procedurally fragile if its score changes when a prestigious affiliation is added or when a confident rebuttal contains no new evidence.

Evidence grounding should also be operationalized rather than treated as a general aspiration. We propose a simple claim-evidence traceability rule for substantive comments that could materially affect scientific evaluation or editorial decisions. Such comments should contain three elements: an evidence anchor identifying the relevant section, paragraph, line range, figure, or table; an evidence type indicating whether the judgment rests on manuscript content, a methodological standard, or external literature; and an uncertainty indicator distinguishing a verified problem from an interpretive concern or uncertain inference. Routine stylistic suggestions need not carry this full burden. The goal is not to make reviews mechanically verbose, but to

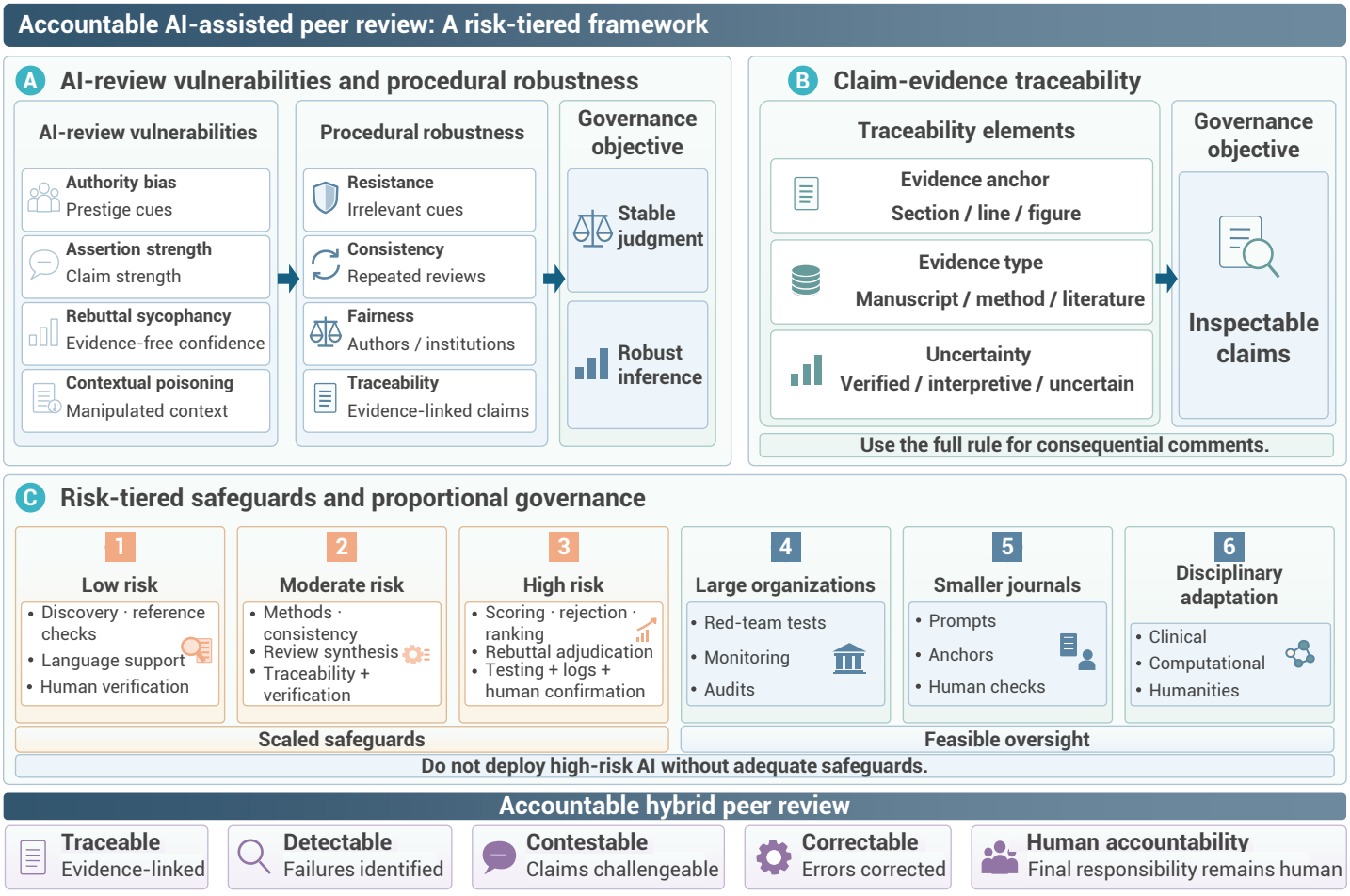


Figure 1. A risk-tiered accountability mechanism for AI-assisted peer review.

ensure that consequential AI-generated criticism can be inspected, challenged, and corrected by editors, reviewers, and authors.

AI-assisted peer review should further be separated by use case because different tasks create different levels of risk. A risk-tiered model is more practical than treating all AI use as a single category. Low-risk assistance includes literature discovery, reference or completeness checks, and language support for reviewer-written comments; these tasks require basic disclosure, confidentiality protection, and human verification. Moderate-risk assistance includes methodological checklist support, consistency checking, identification of potentially missing analyses, and synthesis of multiple human reviews; here, claim-evidence traceability and explicit uncertainty marking should be expected. High-risk uses include automated scoring, desk rejection, acceptance or rejection recommendations, rebuttal adjudication, and manuscript ranking. These decision-bearing tasks warrant the strongest safeguards, including adversarial testing, audit logs, independent human confirmation, and clear assignment of editorial responsibility. Recent work on indirect prompt injection, in which hidden manuscript instructions successfully steer AI-generated reviews, illustrates why high-risk deployment requires security controls beyond prompt engineering alone.[4] Figure 1 summarizes this risk-tiered accountability framework, linking known AI-review vulnerabilities to procedural robustness, evidence traceability, differentiated safeguards, and proportional governance.

As summarized in Figure 1, this tiered approach also addresses implementation cost by matching governance intensity to decision risk and organizational capacity. The intensity of safeguards should be proportional to both decision risk and organizational capacity. Large publishers and major conferences may be able to maintain red-team testing, model-version monitoring, bias audits, and dedicated security infrastructure. Smaller journals, academic societies, and volunteer-led editorial teams may instead rely on standardized prompts, evidence anchoring, mandatory human verification, disclosure rules, and a prohibition on using AI as the final decision-maker. Where high-risk safeguards cannot be supported, high-risk AI use should simply not be deployed. In this sense, accountable governance does not require every journal to build the same technical infrastructure; it requires each organization to limit AI authority to the level it can responsibly supervise.

Disciplinary context should also shape implementation. Clinical and other safety-critical fields may require stronger evidence thresholds and mandatory expert verification because review errors can have downstream consequences for health and safety. Computational fields may place greater emphasis on prompt-injection resistance, reproducibility, and code or benchmark verification. Humanities and qualitative social sciences may require particular care with source attribution, interpretive plurality, and context loss. Procedural robustness should therefore be risk-sensitive rather than discipline-blind, while retaining a common baseline of traceability, human responsibility, and contestability.

Wang et al.[1] demonstrate that the central issue is not whether AI can participate in peer review, but how its participation should be bounded. The realistic objective is not an infallible automated referee. It is a hybrid review system in which AI-induced errors are traceable, detectable, contestable, and correctable before they become editorial decisions. Under such a model, AI may reduce workload and support consistency, while scientific judgment remains embedded in an accountable process in which AI-generated evaluations are traceable, failure modes are detectable, contested judgments are reviewable, errors remain correctable, and final responsibility remains human.

REFERENCES

1. Wang J., Liu Y., Xu H., et al. (2026). When AI reviews science: Can we trust the referee. *Innov. Inform.* **2**:100030. DOI:10.59717/j.xinn-inform.2026.100030
2. Choi B., Jun T.J., Sung J.W., et al. (2026). Invisible text injection and peer review by AI models. *JAMA Netw. Open* **9**:e2552099. DOI:10.1001/jamanetworkopen.2025.52099
3. Vasu S.S.M., Sheth I., Wang H.-P., et al. (2026). Justice in judgment: Unveiling (hidden) bias in LLM-assisted peer reviews. *Find. Assoc. Comput. Linguist. ACL* **2026**:307–330. DOI:10.18653/v1/2026.findings-acl.14
4. Torrielli F., Locci S., Rapp A., et al. (2026). Exploiting large language models in peer review: Indirect prompt injection attacks and integrity probes. *Scientometrics* **131**:4889–4951. DOI:10.1007/s11192-026-05695-x

AUTHOR CONTRIBUTIONS

Y.W. conceptualized the correspondence, drafted the manuscript, revised the text, and approved the final version.

DECLARATION OF INTERESTS

The author declares no competing interests.