When No One Owns the Judgment: Accountability Under Contribution Dissolution in Human–AI Collaboration*

Hengzhi Ye
School of Computer Science
Peking university
Beijing, China
hzye@stu.pku.edu.cn

arXiv:2609.29312v2 [cs.AI] 3 Oct 2026

Abstract

Communities often respond to potentially AI-assisted work by asking three questions: Was AI used? Was that use disclosed? Can hidden use be detected? These questions place AI use itself at the center of accountability while overlooking a deeper problem: unowned judgment. Evaluations, claims, decisions, and creative directions can be shaped by AI with no accountable human or institution prepared to stand behind them. We develop this argument through two illustrative cases: AI-assisted peer review and concealed AI use in creative work. The first shows how contribution dissolution can weaken responsibility while the second shows how the fear of losing credit can discourage honest disclosure. The cases expose the limits of disclosure rules and provenance records as responses to AI-mediated collaboration. We offer three directions for discussion: distinguishing the roles AI plays, identifying judgments that require clear human ownership, and creating conditions in which AI involvement can be disclosed without default penalty. The broader aim is to make AI-shaped contributions discussable, creditable, contestable, and repairable.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Collaborative and social computing**.

Keywords

human-AI collaboration, accountability, contribution dissolution

1 Introduction: Beyond "Did You Use AI?"

Human-AI collaboration rarely follows a clean separation of contribution. A person may write every sentence while the system shapes the central angle, constraints, or evaluative criteria; a system may generate most of the text while simply carrying out a plan already formed by the human. The final outcome therefore reveals little about how the underlying judgments emerged. This reflects a long-standing concern in HCI that human and machine agency is configured through situated interaction instead of being divided along a fixed boundary [15].

Many institutional responses reduce this complex process to a simple question: *Did you use AI?* The question fits easily into a disclosure form, a policy statement, or a detection system, yet it treats different forms of assistance as equivalent. It directs attention to the presence of AI while leaving aside how judgments are formed and who stands behind them. The deeper problem arises when an AI-shaped evaluation, claim, decision, or creative direction enters consequential work without an accountable actor who can answer for its consequences, which we refer to as *unowned judgment*.

The distinction becomes clearer in peer review. A reviewer may use AI to refine the wording of a review they developed independently while retaining full understanding of and responsibility for the underlying judgment. In a different case, a reviewer may submit model-generated objections without being able to evaluate or defend them. Although both practices involve AI, they raise different accountability concerns. A similar distinction applies in journalism: using AI to brainstorm headlines differs substantially from adopting an AI-shaped factual framing that has not been verified by an editor.

Policies centered on AI use struggle to account for these differences. They may penalize limited forms of assistance while failing to specify when a person has delegated a judgment they were expected to make themselves. Research on computerized accountability has long shown that distributed systems can make responsibility difficult to locate [13]. Work on moral crumple zones further shows how responsibility may fall on the nearest human even when that person had little meaningful control [2]. Human–AI collaboration creates a related tension: a person may retain formal responsibility for a judgment whose substance was largely shaped by an AI system.

The core provocation is therefore simple: asking whether AI was used may still matter, but it cannot tell us whether anyone truly owns the resulting judgment. Accountability should begin with the judgments that people and institutions are expected to understand, endorse, and defend. It should also ask who is responsible for correction when those judgments cause harm.

2 Related Work: Contribution Attribution

Contribution attribution becomes difficult when AI shapes more than the final artifact. A system may influence goals, constraints, and intermediate judgments through suggestions, clarifying questions, or generated alternatives. These indirect effects make human and AI contributions difficult to separate from the output alone.

Recent work has therefore turned to the collaborative process itself. Kim et al. [10] propose a goal-level attribution framework that decomposes collaborative goals into requirements and traces direct and indirect influence across dialogue turns. Their findings suggest that AI may have limited influence on high-level direction while playing a larger role in shaping lower-level requirements and micro-decisions that users do not always recognize. Other studies show that people assign credit according to the type, amount, and

*Position paper accepted to the workshop "Nobody Did This": Contribution, Originality, and Accountability in Agent-Mediated Collaboration [7] at the 2026 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '26), Salt Lake City, UT, USA, October 2026.

initiative of AI contribution [3], while systems such as DraftMarks use process traces to make AI involvement in co-writing more visible [14].

These approaches clarify where AI influence enters collaboration, yet attribution alone cannot resolve accountability. A trace cannot determine whether a contribution was legitimate, whether it should reduce professional credit, or who is responsible when an AI-shaped judgment causes harm. Such questions depend on the obligations and norms of the setting. Contribution traces can provide evidence for accountability decisions, but do not resolve them.

3 Case 1: Peer Review

In peer review, reviewers are assigned to assess the contribution of a paper, identify its weaknesses, weigh evidence, and recommend an outcome on behalf of a scholarly community [12]. The review report is only the visible output, thus accountability rests on whether the reviewer understands and can defend the judgments expressed.

Related policies often approach this responsibility through rules about AI use. For example, the ICML 2026 LLM policy compared a strict track that prohibited LLM use in reviewing with a more permissive track that allowed comprehension support and language polishing while limiting the delegation of critical assessment and review writing [5]. ICML later reported using hidden prompts to identify possible violations of the strict policy, followed by the desk rejection of hundreds of submissions whose reviews appeared to violate it [6]. The episode captures the concern that undisclosed AI use may weaken trust in peer review. It also shows the limits of treating detection as the main response. Detecting model involvement says little about which parts of the review were delegated and whether the reviewer could defend them. These concerns are especially serious given the documented limits of LLMs as paper reviewers and the broader expectation that humans remain responsible for AI-assisted scholarly work [17].

The key distinction is whether the reviewer has taken ownership of the judgment. A model-generated criticism may still be acceptable if the reviewer has independently assessed it, agrees with it, and can defend it from their own understanding of the paper. The problem arises when a reviewer makes an objection, novelty claim, score, or recommendation that they cannot evaluate or explain. A contribution trace can show where a judgment originated or how it was shaped, but it cannot show whether the reviewer has genuinely adopted responsibility for it.

Rules centered on AI use remain useful for setting boundaries and supporting disclosure, but limitations exist as they may penalize acceptable support while remaining vague about the delegation of core evaluative work. Knowing that AI was used does not reveal whether the reviewer understood, endorsed, and could defend the judgments in the review. A fuller account of responsibility therefore requires attention to judgment ownership alongside AI use.

This case points toward accountability infrastructure that distinguishes the roles AI plays and identifies where human ownership must be explicit. In peer review, this could mean separating comprehension support and language editing from critique generation, score calibration, and recommendation. More broadly, institutions should identify the judgments for which a person must understand the reasoning, defend the decision, and take responsibility for the correction. This turns attention from exhaustive records of AI use to the points where accountable judgment matters most.

4 Case 2: Creative Work in Journalism

Creative work presents a different accountability problem. In journalism and other communication professions, work is evaluated not only by the quality of the final output, but also by the editorial care, professional judgment, and effort that it reflects [16]. AI assistance can therefore change how work is valued even when the worker remains responsible for the result. Disclosure may be read as evidence of reduced care, skill, or authorship. Perceived effort is known to shape judgments of value [11], while studies of AI-assisted writing have shown related tensions around ownership, authorship, trust, and agency [1, 8, 9].

A recent study of GenAI use in Chinese newsrooms makes this tension concrete [16]. Based on interviews with newsroom managers, editors, and front-line journalists, Xiao et al. found that GenAI was widely used for drafting, summarizing, translation, and language editing, but remained largely private, informal, and disconnected from collaborative workflows. Journalists were often reluctant to discuss these practices with colleagues because AI assistance could be interpreted as laziness, weak editorial rigor, or diminished professional credibility. The result was a gap between what workers did and what they felt safe to acknowledge.

This condition creates what we call a *dark forest of effort*. When AI assistance is useful but disclosure may reduce professional credit, concealment becomes a reasonable response. Workers hide their own processes while seeing only the finished work of others. Organizations then lose the opportunity to develop shared norms around which uses support quality, which delegate editorial judgment, and which should affect credit. Concealment makes it harder to discuss how contributions were produced and where responsibility should lie.

Contribution tracing cannot resolve this tension on its own. Process traces may support coordination and review, but they can also be treated as evidence that a worker performed less "real" work. Under such conditions, workers have reason to use private tools, avoid logged systems, or withhold their practices from colleagues. The value of visibility therefore depends on how AI involvement is interpreted and accounted.

This case points toward accountability infrastructure that makes AI involvement safe to discuss. In such a setting, AI assistance should be distinguished from the delegation of editorial judgment, and credit for work grounded in verification, source interpretation, and final responsibility should be preserved. The value of visibility depends on whether workers can disclose AI involvement without a default credit loss. The broader challenge is to make AI contribution speakable before it becomes punishable.

5 From Attribution to Accountability

The two cases expose complementary failures. In peer review, a reviewer may remain formally responsible for a judgment they cannot explain or defend, while in creative work, a worker may fully understand and stand behind the final product but lose credit once AI assistance becomes visible. Policies centered on disclosure or detection struggle with both situations, as they say little about how AI

involvement should affect responsibility and credit, or who should respond when AI-shaped work goes wrong. Accountability infrastructure thus should make AI-shaped contribution interpretable and discussable before it becomes grounds for penalty.

A practical starting point is to **describe the role of AI** beyond recording the presence. Editing language, making explanations, proposing alternatives, planning a workflow, verifying a claim, and recommending a decision affect collaborative work in different ways. Some roles mainly alter expression, while others shape evaluation, direction, or action. These differences matter because the relevant question is how the model participated in the judgments on which the work depends. Research on contribution attribution similarly shows that credit judgments vary with the type and initiative of AI contribution [3]. Disclosure should preserve these distinctions instead of reducing them to a single category of AI use.

In addition, **judgment ownership checkpoints** are needed. Institutions can identify decisions for which a person must be able to explain the reasoning, endorse the outcome, respond to challenge, and take responsibility for correction. In peer review, these points can include novelty assessment, weakness identification, and the final recommendation. In journalism, they may include factual framing, source interpretation, editorial voice, and publication risk. Such checkpoints do not require a complete record of every prompt or revision. The purpose is to identify moments where someone must be able to say: I understand this judgment, I endorse it, and I am responsible for its consequences.

A further requirement is to **protect disclosure and enable repair**. Workers have little reason to describe AI involvement honestly when disclosure reduces credit by default or creates a presumption of misconduct. Accountability infrastructure therefore needs shared expectations about acceptable AI roles, channels through which assistance can be disclosed without immediate sanction, and procedures for challenging and repairing AI-shaped work. Accountability extends far beyond the assignment of blame, which concerns who must explain an error, revise a decision, and address the potential error. Without these protections, disclosure may reproduce the tensions around ownership, trust, and agency already documented in AI-assisted writing [1, 4, 8].

This provocation does not assume that more visibility is always desirable. Detailed process records can clarify how AI shaped a collaborative outcome, but they may also enable surveillance or simplistic judgments about effort. Attribution remains valuable because it reveals forms of influence that final artifacts cannot show. However, its value depends on the norms through which institutions interpret that evidence. Provenance can establish that AI participated in the process, but the accountability infrastructure must address more complicated questions, including whether that participation is acceptable, how it should affect credit, whether a person owns the resulting judgment, and what should happen when the work is challenged or causes errors.

6 Conclusion

Contribution dissolution is becoming a normal condition of human-AI collaboration. Disclosure and provenance can reveal AI involvement, but cannot determine who understands, endorses, and answers for the resulting judgments. The two cases expose the cost of this gap: responsibility without ownership in peer review and ownership without credit in creative work. Accountability should therefore focus on the roles AI plays, the judgments that require clear human ownership, and the conditions under which AI involvement can be disclosed, challenged, and corrected.

References

[1] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. The AI Ghostwriter Effect: When Users Do Not Perceive Ownership of AI-Generated Text But Self-Declare as Authors. *ACM Transactions on Computer-Human Interaction* 31, 2, Article 25 (2024), 40 pages. doi:10.1145/3637875

[2] Madeleine Clare Elish. 2019. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society* 5 (2019), 40–60. doi:10.17351/ests2019.260

[3] Jessica He, Stephanie Houde, and Justin D. Weisz. 2025. Which Contributions Deserve Credit? Perceptions of Attribution in Human-AI Co-Creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. doi:10.1145/3706598.3713522

[4] Jess Hohenstein and Malte Jung. 2020. AI as a Moral Crumple Zone: The Effects of AI-Mediated Communication on Attribution and Trust. *Computers in Human Behavior* 106 (2020), 106190. doi:10.1016/j.chb.2019.106190

[5] ICML. 2026. LLM Policy. https://icml.cc/Conferences/2026/LLM-Policy. Accessed July 26, 2026.

[6] ICML. 2026. On Violations of LLM Review Policies. https://blog.icml.cc/2026/03/18/on-violations-of-llm-review-policies/. Accessed July 26, 2026.

[7] Kashif Imteyaz, Mohammad Rashidujjaman Rifat, Divya Ramesh, Steven R. Rick, Simo Hosio, Hauke Sandhaus, Advait Sarkar, Christoph Riedl, and Saiph Savage. 2026. "Nobody Did This": Contribution, Originality, and Accountability in Agent-Mediated Collaboration. doi:10.48550/arXiv.2607.26387 arXiv:2607.26387 [cs.CY] Workshop at the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '26).

[8] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 1–13. doi:10.1145/3290605.3300469

[9] Kowe Kadoma, Marianne Aubin Le Quere, Xiyu Jenny Fu, Christin Munsch, Danae Metaxa, and Mor Naaman. 2024. The Role of Inclusion, Control, and Ownership in Workplace AI-Mediated Communication. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Article 1016, 10 pages. doi:10.1145/3613904.3642650

[10] Eunsu Kim, Jessica R. Mindel, Kyungjin Kim, and Sherry Tongshuang Wu. 2026. "I didn't Make the Micro Decisions": Measuring, Inducing, and Exposing Goal-Level AI Contributions in Collaboration. arXiv:2605.21363 [cs.CL] Preprint.

[11] Justin Kruger, Derrick Wirtz, Leaf Van Boven, and T. William Altermatt. 2004. The Effort Heuristic. *Journal of Experimental Social Psychology* 40, 1 (2004), 91–98. doi:10.1016/S0022-1031(03)00065-9

[12] Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology* 64, 1 (2013), 2–17. doi:10.1002/asi.22784

[13] Helen Nissenbaum. 1996. Accountability in a Computerized Society. *Science and Engineering Ethics* 2, 1 (1996), 25–42. doi:10.1007/BF02639315

[14] Momin Naushad Siddiqui, Nikki Nasseri, Adam J. Coscia, Roy Pea, and Hari Subramonyam. 2026. DraftMarks: Enhancing Transparency in Human-AI Co-Writing Through Interactive Skeuomorphic Process Traces. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Article 862, 22 pages. doi:10.1145/3772318.3791109

[15] Lucy A. Suchman. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions.* Cambridge University Press, Cambridge, UK.

[16] Qing Xiao, Qing Hu, Jingjia Xiao, Hancheng Cao, and Hong Shen. 2026. Can GenAI Move from Individual Use to Collaborative Work? Experiences, Challenges, and Opportunities of Coordinating GenAI into Collaborative Newswork. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM, 1515:1–1515:22. doi:10.1145/3772318.3790984

[17] Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 9340–9351. https://aclanthology.org/2024.lrec-main.816/