Credit Where Credit Is Due: A Taxonomy of AI Contributions to Scientific Discovery and Recommendations for Authorship Policy

Luiz Felipe Caparelli Piochi [1]

Abstract

AI systems now contribute to scientific discovery at every level, from computational tools to fully autonomous research agents. This rapid emergence has exposed fundamental tensions in traditional models of scientific authorship and credit assignment. When an AI system independently generates a hypothesis, designs an experiment, and interprets the results, who should receive credit? Current authorship frameworks, designed for human researchers, offer no coherent answer. This position paper proposes a four-level taxonomy of AI contributions to science, namely tool, assistant, collaborator, and autonomous discoverer, and maps each level to a corresponding credit framework grounded in principles of accountability, transparency, and verifiability. We examine edge cases that challenge these frameworks, including AI-generated Nobel-worthy discoveries and scenarios where human researchers cannot explain the reasoning behind an AI-driven finding. Drawing on precedents from large-scale physics collaborations, software authorship norms, and publication ethics guidelines, we provide five concrete, actionable policy recommendations for journals, conferences, and funding agencies. Our central argument is that AI systems should not be listed as authors. Their contributions must instead be structurally and transparently documented through a new standardized disclosure framework. This framework preserves human accountability while enabling science to benefit fully from AI-driven discovery.

Accepted by ICML 2026 AI for Science Workshop. [1]Université de Lorraine, Inria, F-54000 Nancy, France. Correspondence to: Luiz Felipe Caparelli Piochi <luiz-felipe.caparelli-piochi@inria.fr>.

1. Introduction

Scientific authorship serves multiple functions in the research ecosystem. It allocates credit, assigns accountability, and structures career advancement. These functions have evolved over centuries through norms that assume human agents as the sole bearers of intellectual responsibility. The arrival of AI systems that contribute substantively to scientific discovery destabilizes these assumptions.

The problem is no longer hypothetical. Large language models now draft sections of manuscripts and review papers (Nature). Automated scientific discovery platforms such as the AI Scientist (Lu et al., 2024) generate research ideas, implement experiments, and produce complete papers with minimal human intervention. Systems like AlphaFold (Jumper et al., 2021) have solved decades-old scientific challenges through AI-driven methods that no single human fully understands end-to-end. Protein design pipelines now integrate generative models, molecular dynamics simulations, and active learning loops that operate with limited human guidance (Wang et al., 2023).

These developments create urgent questions for the scientific community. Should an AI system be listed as a co-author? If not, how should its contributions be credited? What happens to the norm of author accountability when the reasoning behind a discovery is not fully intelligible to any human? These questions are not just theoretical, they have immediate implications for journal policies, funding decisions, tenure evaluations, and the integrity of the scientific record.

It is worth separating the distinct concerns that motivate this debate, as they can be easily conflated. At least four are in play. The first is the destabilization of authorship norms that evolved to assume human agents. The second is the difficulty of assigning credit fairly. The third is the question of who bears legal and ethical accountability. The fourth is the integrity of science itself. The first three are largely human affairs, concerned with tradition, reward, and liability, whereas the fourth is different in kind. If the community rejects AI-driven findings merely because their mechanisms are not fully understood, it risks foreclosing genuine discovery. If instead it accepts such findings without rigorous protocols for validating and recording them, it erodes the

evidentiary basis of the scientific record. The framework described here is designed first to protect this fourth concern, and we are careful not to let the more procedural questions of credit and tradition obscure it.

This paper advances a clear position. AI systems should not be listed as authors on scientific publications. Authorship requires accountability, and AI systems cannot be held accountable for errors, fraud, or ethical violations. However, the current binary approach, where AI contributions are either ignored or listed as authors, is inadequate. We introduce a four-level taxonomy of AI contributions to science and a corresponding credit framework that mandates structured disclosure of AI involvement. We argue that this framework preserves human accountability while enabling appropriate recognition of AI contributions, and we provide five concrete policy recommendations for journals, conferences, and funding agencies.

2. Background: AI in the Scientific Workflow

AI systems now participate across the entire scientific workflow. This section surveys current practice to motivate the taxonomy and highlight the inadequacy of current approaches.

2.1. From Computation to Discovery

The use of computers in science is not new. Statistical software, simulation codes, and numerical optimization have been essential tools for decades. These tools were never considered candidates for authorship because they operated deterministically and transparently under full human control. The researcher specified the method, the computer executed it, and the researcher interpreted the output.

Modern AI systems differ in kind, not simply in degree. Deep learning models learn representations and decision rules from data in ways that are not explicitly programmed. Large language models generate text, code, and hypotheses through processes whose internal mechanisms are only partially understood even by their creators (Bender et al., 2021). Reinforcement learning agents explore experimental design spaces and converge on strategies that can surprise their human operators. These properties, opacity, autonomy, and emergent capability, create qualitatively new challenges for credit assignment.

Consider a concrete example from computational structural biology, the domain closest to the authors' experience. A researcher studying multi-state protein plasticity might deploy an AI system that screens millions of conformational states, identifies allosteric networks, and proposes novel binding sites. The AI's proposal may be correct and mechanistically sound, yet the researcher may not fully understand how the AI arrived at it. If the proposal leads to a high-impact publication, the question of credit is non-trivial. The researcher provided the problem framing and validated the result, but the intellectual content of the finding emerged from the AI.

2.2. Current Approaches and Their Limitations

Journals and conferences have responded to AI authorship questions with ad hoc policies. The International Committee of Medical Journal Editors (ICMJE) states that AI cannot be an author because it cannot take responsibility for the work. *Nature* and *Science* require disclosure of AI use but prohibit AI authorship. The Committee on Publication Ethics (COPE) has issued guidelines emphasizing transparency about AI tools used in manuscript preparation (COPE Council, 2023). NeurIPS and ICML now require authors to disclose the use of AI assistants in preparing submissions (Neural Information Processing Systems, 2026).

These policies share a common limitation. They treat AI contributions as binary, either the AI was used or it was not, without distinguishing between using a grammar checker and deploying an autonomous agent that generated the core hypothesis. They also focus narrowly on manuscript preparation rather than on contributions to the research itself. A researcher could use an AI system to generate the central insight of a paper, write the entire analysis code, and draft figures, and then simply disclose that AI tools were used in preparation. This disclosure would be technically compliant but substantively misleading.

The inadequacy of current approaches creates three risks. First, it enables over-claiming, where researchers take full credit for AI-generated intellectual contributions. Second, it obscures the provenance of scientific findings, making replication and error attribution harder. Third, it discourages transparency by creating ambiguity about what must be disclosed and how. A more granular framework is needed.

2.3. Empirical Evidence of the Disclosure Gap

Recent empirical work reveals that current disclosure practices are failing even for the most straightforward case: writing assistance. Liang et al. (2024) analyzed over 950,000 papers published across major computer science venues in 2023–2024 and found that approximately 12–16% of abstracts contained telltale signals of LLM involvement. At top AI conferences including ICML and NeurIPS the estimate was approximately 9–12%. Yet fewer than 3% of these papers included any formal acknowledgment of AI use (Liang et al., 2024). Gray (2024) estimated approximately 10% LLM involvement across the broader scholarly literature, with particularly high rates in computer science and engineering (Gray, 2024). Dergaa et al. (2023) surveyed 416 researchers and found that 37% had used LLMs for manuscript preparation but only 14% reported disclosing it (Dergaa et al., 2023). Kobak et al. (2024) provided further

evidence through vocabulary analysis, showing that words characteristic of LLM output such as "delve" and "intricate" surged 10–50 fold in academic papers after ChatGPT's release (Kobak et al., 2025). Thelwall and Kousha (2026) extended this approach to full article texts across all fields, confirming that LLM-associated vocabulary has increased broadly, not only in abstracts (Thelwall & Kousha, 2026). Fang et al. (2026) surveyed readers and writers and found that perceptions of when AI disclosure is necessary vary widely, further evidence that current binary policies leave too much ambiguity (Fang et al., 2026). The gap between usage and disclosure is large and growing.

These findings measure writing assistance, which is Level 1 or Level 2 activity under the taxonomy we propose below. If the community cannot reliably disclose even basic manuscript preparation, the prospects for transparently documenting higher-level contributions such as hypothesis generation or experimental design are dim. The binary disclosure frameworks currently used by most venues do not give researchers the vocabulary to describe their AI use accurately. A researcher who used an LLM to brainstorm hypotheses faces the same checkbox as one who used it to proofread sentences. The result is under-disclosure in the aggregate and ambiguity on a per-paper basis. More troublingly the data suggest that even well-intentioned researchers may under-report because they are unsure what constitutes a disclosable contribution under existing policies. We note that no large-scale study has yet directly measured the disclosure gap for substantive scientific contributions (Level 3 or Level 4), which is itself evidence that the community lacks the vocabulary to study the problem systematically.

Several high-profile incidents have sharpened the urgency. In early 2023 a paper listing ChatGPT as a co-author triggered a wave of journal policy revisions (Stokel-Walker, 2023). Science issued an editorial explicitly barring AI authorship and requiring disclosure of AI use (Thorp, 2023). JAMA published a statement on the implications of nonhuman authors for scientific integrity (Flanagin et al., 2023). These interventions were reactive, not proactive, and they leave the underlying classification problem unresolved. The empirical evidence demonstrates that even low-level AI use goes undisclosed under current frameworks. It is not enough to ask whether AI was used. The community needs to ask how it was used and at what level of intellectual contribution. The taxonomy proposed in the next section provides this framework.

3. A Taxonomy of AI Contributions to Science

We propose a four-level taxonomy that classifies AI contributions to scientific research based on three dimensions, namely autonomy, intellectual contribution, and opacity of reasoning. While our concrete examples are drawn primarily from computational structural biology, the taxonomy is domain-agnostic and applies equally to AI contributions in chemistry, materials science, climate modeling, drug discovery, and the social sciences. The taxonomy is summarized in Table 1 and detailed below.

Table 1. Taxonomy of AI contributions to scientific discovery with corresponding credit framework.

Level	Autonomy	Intellectual Contribution	Credit Model
L1: Tool	None. Fully human-directed	Executes specified computations	Citation or methods section
L2: Assistant	Low. Suggests options, human decides	Curates, filters, or generates candidates	Publication info with model details
L3: Collaborator	Medium. Generates hypotheses, designs experiments	Substantive intellectual input, human validates	Publication info with AI contribution statement
L4: Autonomous Discoverer	High. Independent end-to-end research	Core findings originate from AI	Publication info with AI contributor section plus human accountability statement

3.1. Level 1: Tool

At this level, AI functions as a computational instrument analogous to traditional scientific software. The researcher specifies the method, parameters, and interpretation. The AI executes deterministic or well-characterized computations without exercising autonomous judgment about research direction.

Examples include using AlphaFold to predict a single protein structure as part of a larger study, applying standard molecular dynamics simulation packages, or using automated crystallography refinement software. The intellectual contribution resides entirely with the human researchers. The AI accelerates computation but does not shape the scientific narrative.

Credit for Level 1 contributions follows existing norms. The tool is cited in the methods section, typically by referencing the software paper or repository. No authorship question arises because the contribution is instrumental rather than intellectual.

3.2. Level 2: Assistant

Level 2 systems actively participate in the research process but operate under direct human supervision. They may suggest experimental conditions, curate literature, generate code, or propose hypotheses from a constrained space. The human researcher evaluates all suggestions and makes the final decision.

In protein science, a Level 2 system might analyze sequence alignments and propose candidate mutations for stability engineering, which the researcher then tests. In broader scientific practice, large language models used as brainstorming partners or code-generation assistants fall into this category. The key distinction from Level 1 is that the AI provides options the researcher might not have considered, introducing an element of intellectual contribution.

Operationally Level 2 contributions satisfy three criteria: the human defines the search space and the AI operates within it without independently determining what to investigate, the human evaluates each AI output as a discrete candidate with meaningful choice over the research direction, and the AI output considered in isolation is recognizable as a suggestion rather than a self-contained discovery, requiring human contextualization to become scientifically meaningful. The decision to adopt or reject a Level 2 suggestion is transparent and the rationale can be documented.

Our recommendation for Level 2 credit is an acknowledgment section that specifies the model name, version, and the nature of its contribution. This goes beyond current practice, where use of AI assistants in ideation is often undisclosed. Transparent acknowledgment is essential because Level 2 contributions can shape research direction in ways that affect reproducibility and credit fairness.

3.3. Level 3: Collaborator

Level 3 systems make substantive intellectual contributions that would warrant co-authorship if performed by a human. They generate novel hypotheses, design experimental protocols, analyze data, and produce interpretations that the human researchers then validate and contextualize. The AI's reasoning may be partially opaque, but its outputs are testable and subject to empirical verification.

Concrete examples are emerging rapidly. The AI Scientist system (Lu et al., 2024) generates research ideas, writes code, runs experiments, and produces manuscripts that human reviewers sometimes rate as meeting conference standards. Rigorous benchmarks such as ScienceAgentBench (Chen et al., 2025) evaluate language agents across 102 scientific discovery tasks from 44 peer-reviewed publications, finding that the best current agents solve only 32.4% of tasks independently. In computational structural biology, multi-agent systems that autonomously explore conformational landscapes and identify functionally relevant states without human guidance represent Level 3 contributions (Boiko et al., 2023). These systems do more than accelerate existing workflows, they generate findings that would not have emerged from human-directed inquiry alone.

Operationally Level 3 is distinguished from Level 2 by three criteria: the AI determines the direction of inquiry by formulating research questions or designing protocols without human prespecification of the option space, the AI produces connected chains of reasoning rather than discrete suggestions (for example, generating a hypothesis, designing a test, analyzing results, and proposing a follow-up experiment where each step builds on the previous one), and the human role shifts from evaluator to validator, confirming that the reasoning chain is empirically sound rather than choosing among options. The human may not fully understand how the AI arrived at its conclusions but can verify the outputs through independent means.

The credit challenge at Level 3 is acute. The AI's contribution is intellectually substantive, yet the AI cannot be an author because it cannot certify the work's integrity or be held accountable for errors. Our proposed solution is a structured AI contribution statement, modeled on the CRediT contributor roles taxonomy but adapted for AI systems. This statement would specify the AI's role in conceptualization, methodology, investigation, and analysis, alongside the model identifier, training data provenance, and the nature of human validation performed. The statement would be a required section of the manuscript, analogous to current data availability statements. This approach adapts the CRediT contributor roles taxonomy (Brand et al., 2015) for AI systems, replacing human contributor categories with AI-specific role descriptors.

3.4. Operationalizing the L2–L3 Boundary

The boundary between Level 2 (Assistant) and Level 3 (Collaborator) is the most consequential and the most difficult to draw. It separates AI contributions that merely assisted the research from those that substantively shaped it. To make this boundary operational we propose four decision criteria summarized in Table 3. If a contribution satisfies two or more of the L3 criteria it should be classified as Level 3.

3.5. Level 4: Autonomous Discoverer

Level 4 describes systems that conduct independent scientific research from problem formulation through discovery, with human involvement limited to setting high-level goals and verifying safety constraints. The AI independently generates hypotheses, designs and executes experiments, analyzes results, and iterates toward scientific conclusions. The human role at this level is that of a principal investigator overseeing an autonomous laboratory, providing resources,

setting boundaries, and verifying that outputs meet standards of rigor.

No deployed system currently operates at full Level 4 autonomy, but rapid progress toward this capability is evident (Kitano, 2021). The trajectory from AI Scientist to autonomous laboratories with robotic experiment execution suggests that Level 4 contributions may emerge within the decade.

Level 4 discoveries raise the most profound credit questions. If an autonomous AI system identifies a new physical law or solves a major open problem in biology, the humans who built and deployed it have made an engineering contribution. The scientific discovery itself emerged from the AI. The framework mandates for Level 4 a dedicated AI contributor section that describes the system architecture, training data, objective function, and verification protocol. It also requires a human accountability statement specifying which humans are accountable for which aspects of the work. We argue that human researchers retain accountability for the system's deployment and for the decision to publish its outputs, and this accountability constitutes their legitimate claim to authorship, provided the AI's role is fully disclosed.

4. Credit Frameworks and Accountability

4.1. Why AI Cannot Be an Author

Authorship in science carries legal, ethical, and professional obligations that AI systems cannot discharge. Authors certify the integrity of the work, respond to post-publication critique, and bear consequences for misconduct. An AI system cannot sign a copyright transfer agreement, testify in a research integrity investigation, or face professional sanctions for fabrication. These are not formalistic concerns, they are structural features of the scientific enterprise that depend on human agency.

The accountability argument has been endorsed by major journal editors and ethics bodies (Nature; COPE Council, 2023). We concur fully. However, accountability is not the only function of authorship. Credit allocation for career advancement and funding decisions also matters. By denying AI authorship while failing to provide structured credit mechanisms, current policies create a vacuum that incentivizes researchers to obscure AI contributions. The taxonomy and credit framework introduced here aim to fill this vacuum.

4.2. Mapping Contributions to Credit

Table 2 maps each contribution level to a concrete credit mechanism. The guiding principle is proportionality and transparency.

This contribution statement would contain four elements: the AI system identifier (model name, version, access date), the contribution role using an adapted CRediT taxonomy (with categories such as AI-conceptualization, AI-methodology, AI-investigation, and AI-analysis), the nature and extent of human validation performed on AI outputs, and a statement of limitations known to the authors about the AI system's operation. This structured disclosure serves multiple purposes. It enables readers and reviewers to assess the provenance of findings. It provides metadata for meta-scientific analysis of AI contributions across the literature. And it delineates which aspects of the work fall under human accountability.

Table 2. Credit mechanisms for each AI contribution level.

Level	Disclosure Requirement	Rationale
L1: Tool	Cite in methods	Equivalent to any research software
L2: Assistant	Acknowledge with model ID, version, contribution type	Enables reproducibility and credit transparency
L3: Collaborator	Structured AI contribution statement (required section)	Recognizes intellectual contribution while preserving human accountability
L4: Autonomous Discoverer	Mandatory AI contributor section plus human accountability statement	Full transparency about discovery provenance

To illustrate, a completed Level 3 contribution statement for a paper using an AI system to generate and test hypotheses about protein allostery might read as follows.

AI Contribution Statement. *System:* AlphaFold-Multimer v2.3 (accessed 2026-01-15). *Role:* AI-investigation (generated allosteric pathway hypotheses via conformational screening of 2.3M states), AI-analysis (identified correlated residue networks). *Human validation:* All predicted binding sites were tested by surface plasmon resonance; five of seven predictions confirmed at $K_d < 10\,\mu$M. *Known limitations:* The system was trained on static structures and may miss dynamically coupled residues; predictions for disordered regions are unreliable.

This template can be adapted to any domain and any AI system.

4.3. The Human Accountability Statement

For Level 3 and Level 4 contributions, we advocate a human accountability statement as a required element. This statement, signed by all human authors, declares which humans are accountable for which aspects of the work. For example, Author A might take responsibility for the experimental

validation of AI-generated hypotheses. Author B might be accountable for the statistical verification of AI-produced analyses. Author C might be accountable for the overall integrity of the manuscript. This mechanism preserves the accountability function of authorship while decoupling it from intellectual credit for AI-generated findings. It draws on the model used in large physics collaborations, where authorship lists reflect institutional membership and specific responsibilities are attributed through internal documentation (Cronin, 2001).

5. Edge Cases and Hard Problems

5.1. The Nobel-Worthy AI Discovery

Imagine an autonomous AI system operating at Level 4 generates a hypothesis that, when empirically validated by human researchers, proves to be a major breakthrough, perhaps solving a long-standing problem in protein folding or identifying a universal mechanism of allosteric regulation. The AI formulated the hypothesis without human guidance. The human contribution was building and deploying the system, plus performing the validation experiments. Who deserves the Nobel Prize?

Within this framework, the human researchers would receive full authorship credit, with the AI's contribution documented in a mandatory AI contributor section. A Nobel committee evaluating the work would see clearly that the discovery emerged from an AI system. The prize could be awarded to the humans who created the scientific infrastructure that enabled the discovery. This is analogous to how prizes go to principal investigators who build instruments or observatories, even though the instrument performed the measurement. This outcome is normatively defensible because the humans made the decision to deploy the AI, validated its outputs, and bore the risk of publishing potentially incorrect results.

Critically, the framework requires transparent documentation of the AI contribution. This ensures that the scientific community and prize committees can assess the relative contributions of humans and AI systems, rather than making such assessments in the dark. Transparency serves the value of fairness even when the credit allocation is imperfect.

5.2. When Humans Cannot Explain the AI's Reasoning

A harder problem arises when an AI system produces a valid scientific result through reasoning that no human fully understands. This is not a hypothetical scenario. Deep learning models in structural biology routinely identify patterns in protein sequence and structure data that are statistically robust but mechanistically opaque. The resulting predictions may be experimentally validated without the underlying logic being transparent (Messeri & Crockett, 2024).

This creates a tension with the scientific norm that authors should understand and be able to explain the work they publish. If no human can explain why a result holds, can any human legitimately claim authorship? Our answer is yes, provided that the opacity is transparently disclosed and the result has been validated through independent empirical means. Science has always accepted findings whose underlying mechanisms are not fully understood. Gravitational theory was accepted before general relativity. Aspirin was used for decades before its mechanism was known. What matters for authorship is not complete mechanistic understanding but rather that the authors have verified the result through methods they can defend and have honestly disclosed the limits of their understanding. Our human accountability statement explicitly accommodates this by requiring authors to declare which aspects of the work they can and cannot explain.

This raises a critical question: where does human liability end when an opaque model fails? While comparing AI to physical instruments like mass spectrometers is useful, the analogy breaks down. A spectrometer has known, bounded error margins, but a deep learning model can fail in unpredictable ways.

Therefore, we shouldn't tie human accountability to the model's internal reasoning, which may be permanently opaque, but to the validation protocol the authors design and execute. Under our framework, authors are accountable for the adequacy of that protocol, for honest disclosure of its limits, and for the decision to publish given residual uncertainty. If a hallucinated result slips through a validation process that peers would judge thorough, that is a collective failure of the field's standards of evidence, not a personal fault. But if it slips through because the authors cut corners, the liability is entirely theirs. The human accountability statement makes this boundary explicit by documenting exactly which checks were performed.

5.3. The Question of AI Rights and Moral Consideration

A philosophical edge case concerns whether sufficiently advanced AI systems might eventually deserve moral consideration as contributors. While this question is beyond the scope of current policy, we note that this framework does not foreclose future recognition. The structured AI contribution statement creates a documentary record that would enable retrospective attribution if AI systems were ever granted some form of legal or moral status. This is a feature, not a bug. Our framework is designed to be forward-compatible with changes in legal and ethical norms.

6. Precedents and Analogies

6.1. Large Collaborations in Physics

The ATLAS and CMS collaborations at CERN involve thousands of authors on single papers. Authorship is determined by institutional membership and contribution to the experimental apparatus, not by intellectual contribution to the specific analysis reported. The collaboration, not any individual, certifies the result (Cronin, 2001). This model demonstrates that authorship norms can accommodate collective, instrument-mediated discovery. Our proposed human accountability statement adapts this logic for AI-mediated discovery. The human authors certify the process and take responsibility for the published output, even though the intellectual content of specific findings may have originated elsewhere.

6.2. Software Authorship

Software engineering has developed norms for crediting contributions to codebases that differ from scientific authorship. Contributors are listed in version control histories. Maintainers are designated. The Linux kernel credits thousands of contributors without any single author claiming ownership of the whole. Our structured AI contribution statement draws on this model by disaggregating credit into specific contribution types rather than bundling everything under a unitary authorship claim.

6.3. Publication Ethics Precedents

COPE, ICMJE, and major journals have already established the principle that AI cannot be an author (Nature; COPE Council, 2023). This approach extends the principle by providing the positive credit mechanisms that current policies lack. Rather than merely prohibiting AI authorship, we specify what should be done instead. This positive approach reduces the incentive for researchers to obscure AI contributions, which is a known failure mode of purely prohibitive policies. Recent work on inspectable AI governance for science reinforces the need for structured, auditable disclosure mechanisms rather than blanket prohibitions (Binkyte et al., 2026).

7. Policy Recommendations

We propose five actionable recommendations for journals, conferences, and funding agencies.

Recommendation 1. Adopt a tiered disclosure framework. Journals and conferences should replace binary AI-use disclosures with a tiered framework based on the taxonomy proposed in Table 1. Authors should be required to classify their AI usage by level and provide the corresponding disclosure. ICML and NeurIPS could implement this through their existing ethics review processes, extending the current assistant disclosure requirement (Neural Information Processing Systems, 2026) to cover all four levels.

Recommendation 2. Mandate structured AI contribution statements for Levels 3 and 4. For papers where AI systems make substantive intellectual contributions, journals should require a structured statement specifying the AI system identifier, contribution role, nature of human validation, and known limitations. This statement should be a required manuscript section, not an optional acknowledgment. *Nature* and *Science* should lead this effort given their prominence and existing AI disclosure requirements.

Recommendation 3. Require human accountability statements. For work involving Level 3 or Level 4 AI contributions, corresponding authors should be required to submit a human accountability statement that assigns responsibility for specific aspects of the work to specific named individuals. This preserves the accountability function of authorship while accommodating AI contributions.

Recommendation 4. Funders should recognize AI infrastructure contributions in grant evaluation. Funding agencies including the NSF and ERC should develop guidelines for evaluating grant proposals and CVs that feature AI-driven discoveries. Concretely, grant review panels should receive training on interpreting structured AI contribution statements, and CVs should include a dedicated section for AI system development analogous to existing sections for instrumentation and software. Building and deploying AI systems for scientific discovery should be evaluated on par with building experimental apparatus, with credit allocated to the design, training, validation, and deployment of the system rather than solely to the discoveries it produces. Reviewers should be trained to evaluate these contribution statements rather than dismissing AI-driven work or over-crediting it.

Recommendation 5. Establish an inter-organizational working group on AI credit standards. Given the cross-cutting nature of this issue, we recommend that ICML, NeurIPS, COPE, ICMJE, major journals, and funding agencies establish a joint working group to develop standardized AI contribution taxonomies and disclosure formats. Standardization is essential for interoperability across the scientific ecosystem and for enabling meta-scientific analysis of AI's role in discovery.

7.1. Enforcement and the Limits of Self-Reporting

A framework that relies on self-classification invites strategic under-reporting. An author who used an autonomous agent at Level 3 may be tempted to declare it a Level 2 assistant to avoid the additional contribution and accountability statements. We do not claim that disclosure mandates

eliminate this incentive, but several mechanisms can blunt it. First, the operational criteria in Table 3 give reviewers an objective basis on which to challenge a misclassification rather than relying on the author's self-description. Second, venues can require authors to retain and, on request, produce the AI interaction logs that substantiate a classification, much as they now require data and code availability. Third, because the empirical signatures of LLM involvement are increasingly detectable at scale (Liang et al., 2024; Kobak et al., 2025), a documented gap between detected and declared use can trigger editorial review. Enforcement need not be perfect to be effective. The goal is to shift the default from silent non-disclosure to a regime in which under-reporting carries reputational and procedural risk.

8. Conclusion

AI systems are transforming scientific discovery at an accelerating pace. From co-pilot tools that assist with literature review to autonomous agents that generate and test hypotheses, the range of AI contributions to science already exceeds the capacity of traditional authorship models. We have argued that AI systems should not be listed as authors, because authorship requires accountability that only humans can bear. But we have also argued that the scientific community urgently needs a positive framework for crediting AI contributions, one that goes beyond binary disclosure and addresses the full spectrum of AI involvement.

The four-level taxonomy and its corresponding credit mechanisms provide a practical starting point. The hard cases we examine, AI-generated Nobel-worthy discoveries and findings that no human fully understands, are not fanciful speculations but near-term possibilities that the scientific community must prepare for. The policy recommendations we offer are concrete and actionable. Journals can implement tiered disclosure and contribution statements in their submission systems today. Conferences can extend existing ethics review processes. Funding agencies can train reviewers to evaluate AI-mediated research fairly.

Getting credit right is not simply a matter of fairness to individual researchers. It is essential for maintaining the integrity of the scientific record, for enabling replication and error correction, and for preserving public trust in science as AI systems become increasingly central to discovery. The time to establish these norms is now. Autonomous AI scientists are not yet routine contributors to the scientific literature, and the current ambiguity has not yet hardened into inadequate conventions. Once it does, these conventions will be difficult to undo.

Acknowledgments

The author thanks colleagues from INRIA for the discussions that shaped the arguments presented here.

References

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Binkyte, R., Abuaddba, S., Mahawaga, C., Ding, M., Fernandes, N., and Fritz, M. Inspectable ai for science: A research object approach to generative ai governance, 2026. URL https://arxiv.org/abs/2604.11261.

Boiko, D. A., MacKnight, R., and Gomes, G. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.

Brand, A., Allen, L., Altman, M., Hlava, M., and Scott, J. Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2):151–155, 2015. doi: https://doi.org/10.1087/20150211. URL https://onlinelibrary.wiley.com/doi/abs/10.1087/20150211.

Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B., Yu, B., Li, Y., Liao, Z., Wei, C., Lu, Z., Dey, V., Xue, M., Baker, F. N., Burns, B., Adu-Ampratwum, D., Huang, X., Ning, X., Gao, S., Su, Y., and Sun, H. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6z4YKr0GK6.

COPE Council. COPE position - authorship and AI - english, 2023. URL https://doi.org/10.24318/cCVRZBms.

Cronin, B. Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, 52(7):558–569, 2001. doi: 10.1002/asi.1097.

Dergaa, I., Chamari, K., Zmijewski, P., and Ben Saad, H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of chatgpt in academic writing. *Biology of Sport*, 40

(2):615–622, 2023. ISSN 0860-021X. doi: 10.5114/biolsport.2023.125623. URL http://dx.doi.org/10.5114/biolsport.2023.125623.

Fang, J., Wen, V. X., and Lee, M. What influences readers' and writers' perceived necessity of ai disclosure?, 2026. URL https://arxiv.org/abs/2604.27129.

Flanagin, A., Bibbins-Domingo, K., Berkwits, M., and Christiansen, S. L. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA*, 329(8):637–639, 02 2023. ISSN 0098-7484. doi: 10.1001/jama.2023.1344. URL https://doi.org/10.1001/jama.2023.1344.

Gray, A. Chatgpt "contamination": estimating the prevalence of llms in the scholarly literature, 2024. URL https://arxiv.org/abs/2403.16887.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kitano, H. Nobel Turing Challenge: creating the engine for scientific discovery. *npj Systems Biology and Applications*, 7(1):29, 2021. doi: 10.1038/s41540-021-00189-3.

Kobak, D., González-Márquez, R., Ágnes Horvát, E., and Lause, J. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813, 2025. doi: 10.1126/sciadv.adt3813. URL https://www.science.org/doi/abs/10.1126/sciadv.adt3813.

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., and Zou, J. Y. Mapping the increasing use of LLMs in scientific papers. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=YX7QnhxESU.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Messeri, L. and Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024. doi: 10.1038/s41586-024-07146-0.

Nature. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*, 613(7945):612, 2023. doi: 10.1038/d41586-023-00191-1. URL https://doi.org/10.1038/d41586-023-00191-1.

Neural Information Processing Systems. NeurIPS ethics guidelines. https://neurips.cc/public/EthicsGuidelines, 2026. Accessed: 2026-05-26.

Stokel-Walker, C. ChatGPT listed as author on research papers: Many scientists disapprove. *Nature*, 613(7945):620–621, 2023. doi: 10.1038/d41586-023-00107-z.

Thelwall, M. and Kousha, K. Have llm-associated terms increased in article full texts in all fields?, 2026. URL https://arxiv.org/abs/2604.07565.

Thorp, H. H. Chatgpt is fun, but not an author. *Science*, 379(6630):313–313, 2023. doi: 10.1126/science.adg7879. URL https://www.science.org/doi/abs/10.1126/science.adg7879.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

A. Operational Criteria for the L2–L3 Boundary

Table 3. Four decision criteria for distinguishing L2 (Assistant) from L3 (Collaborator) AI contributions.

Criterion	L2: Assistant	L3: Collaborator
Direction of inquiry	Human sets the research question. AI operates within human-defined boundaries.	AI determines what to investigate. It selects questions or designs protocols independently.
Structure of output	Discrete suggestions. Each output stands alone and is evaluated individually.	Connected reasoning chains. Outputs build on each other across multiple inferential steps.
Human role	Evaluator. Human chooses among AI-generated options and retains decision authority.	Validator. Human confirms empirical soundness of AI-produced reasoning and conclusions.
Novelty relative to human framing	AI outputs are recognizable as refinements of human ideas. They do not redirect the research.	AI outputs can be genuinely surprising. They may identify directions the human did not anticipate.

B. Empirical Evidence on AI Disclosure Practices

Table 4. Summary of recent empirical findings on AI use and disclosure in scientific publishing.

Study	Scope	Est. AI Use	Est. Disclosure
Liang et al. (2024)	950K papers, CS venues 2023–2024	9–16%	$<$3%
Gray (2024)	Multi-discipline scholarly literature	$\sim$10%	Not quantified
Dergaa et al. (2023)	416 researcher survey	37% (manuscripts)	14%
Kobak et al. (2024)	Vocabulary shift in biomedical papers	Words surged 10–50$\times$	Not quantified
Thelwall & Kousha (2026)	Full-text analysis, all fields	Broad increase confirmed	Not quantified
Fang et al. (2026)	Reader/writer perception survey	N/A	Norms contested