THE VERIFICATION BOTTLENECK: MANAGING TRUST IN POST-AGI SCIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

AI systems can now compress decade-long research programs into days. The result is a structural asymmetry: discovery scales faster than verification. Current AI achieves only 21% recall detecting errors in manuscripts; it generates plausible claims far better than it verifies them. This "verification bottleneck" has been flagged as a concern, but we lack frameworks for managing it. We contribute three operationalized mechanisms: (1) *epistemic triage* combining prediction markets, statistical thresholds, and anomaly detection to prioritize what gets verified; (2) *verification cascades*, a hierarchical architecture assigning epistemic status based on verification depth; and (3) an extension of *provisional knowledge* to AI-generated claims, with explicit conditions for status transitions. Drawing on social epistemology and the sociology of scientific knowledge, we examine how these frameworks address scalable oversight, trust, and human roles in machine-accelerated science. Without such frameworks, science risks epistemic pollution: a state where valid and invalid claims become indistinguishable.

1 INTRODUCTION

Will machines generate discoveries for humans to validate, or will humans generate questions for machines to resolve? This framing misses the more urgent problem: what happens when machines generate discoveries faster than any verification infrastructure (human or automated) can validate them?

The numbers are sobering. Google DeepMind's AI co-scientist compressed what researchers estimated as a decade of hypothesis development into two days, identifying novel therapeutic targets for antimicrobial resistance (Gottweis et al., 2025). LLM-assisted researchers publish roughly 33–50% more papers than non-users, with larger gains for non-native English speakers (Kusumegi et al., 2025). Scientific literature already doubles every nine years (Bornmann & Mutz, 2015). Yet AI systems achieve only 21% recall and 6% precision detecting errors in manuscripts (Son et al., 2025). AI is far better at generating plausible claims than verifying them.

This "verification bottleneck," the structural mismatch between generation and verification, has been identified as an emerging challenge (Cornelio et al., 2025). Post-AGI, this bottleneck becomes absolute: if AGI systems generate discoveries requiring years of human verification each, and produce thousands per hour, no verification infrastructure can keep pace. The question is not whether unverified AI-generated knowledge will circulate in science. It is how we manage a corpus where verified knowledge becomes a shrinking fraction of total claims. What's missing is a framework for operating within this reality. How do we maintain oversight when claims proliferate faster than they can be checked? How do we preserve trust when verification becomes impossible for most claims? What role remains for humans when discovery is automated but verification can't be?

We develop three interconnected frameworks. First, we operationalize *epistemic triage* (Kelly, 2025) with concrete mechanisms for prioritizing verification. Second, we introduce *verification cascades* as a hierarchical validation architecture. Third, we extend *provisional knowledge* (Teller, 2009) to AI-generated claims with explicit status transition conditions. The epistemology of testimony (Hardwig, 1991; Lackey, 2008) and sociology of scientific knowledge (Shapin, 1994) ground these frameworks theoretically.

2 THE VERIFICATION BOTTLENECK

2.1 QUANTIFYING THE ASYMMETRY

Three trends converge to create the bottleneck: accelerating production, strained verification infrastructure, and asymmetric AI capabilities.

Production acceleration. Around 2.5 million peer-reviewed papers appear annually, with volume doubling every nine years (Bornmann & Mutz, 2015). AI introduces qualitative acceleration on top of this already-strained system. Sakana AI's "AI Scientist" autonomously generates papers that, according to developers, have "passed peer review" (Lu et al., 2024). Independent evaluation tells a different story: 42% experiment failure rate, acceptance only at workshops with 60–70% acceptance rates (Beel et al., 2025). "Passed peer review" and "verified knowledge" aren't the same thing.

Verification constraints. Peer review is buckling. Twenty percent of researchers perform 69–94% of all reviews; editors must invite 10–12 reviewers to secure two; the global annual cost exceeds $2.7 billion (Kovanis et al., 2016). Replication faces worse odds: $52,000–$75,000 per paper, and only 36% of psychology studies replicate successfully (Open Science Collaboration, 2015). Ioannidis (2005) argued that most published findings are false even before AI entered the picture. Bias, low power, and analytic flexibility saw to that. Nonreplicable papers get cited more than replicable ones (Serra-Garcia & Gneezy, 2021).

AI asymmetry. Could AI verify AI-generated claims, closing the loop? Current evidence says no. The SPOT benchmark finds state-of-the-art models achieve under 21% recall and 6% precision detecting planted errors (Son et al., 2025). The asymmetry runs deep: generating plausible claims requires only local coherence; verification demands global consistency with all relevant evidence.

2.2 THE EPISTEMOLOGICAL STAKES

This isn't just a scaling problem; it threatens the epistemic foundations of science. Hardwig (1991) showed that modern science already depends extensively on trust: researchers accept testimony from specialists without independent verification. But Hardwig's framework assumed human agents with reputations, ethical commitments, accountability. AI systems traditionally satisfy none of these.

Shapin (1994) demonstrated that scientific knowledge is constituted through social processes of credibility assessment. Koskinen (2024) argues we lack a satisfactory social epistemology for AI-based science. If AI-generated knowledge can't be integrated into credibility assessment, if we must simply accept or reject it without engaging the underlying reasoning, we risk splitting knowledge into human-verified (socially constituted) and AI-generated (merely accepted) categories.

3 EPISTEMIC FRAMEWORKS FOR THE VERIFICATION BOTTLENECK

We propose three interconnected frameworks. The concepts themselves have precedents in epistemology; our contribution is operationalizing and integrating them for AI-generated science.

3.1 OPERATIONALIZING EPISTEMIC TRIAGE

Epistemic triage, scaling verification effort to stakes, has been proposed for AI knowledge claims (Kelly, 2025). We contribute three mechanisms for making it concrete:

Statistical thresholds. Stricter significance criteria ($p < .005$ rather than $p < .05$) automatically prioritize claims more likely to be true (Benjamin et al., 2018). Low-cost filtering, no human judgment required.

Prediction markets. Let researchers bet on replication outcomes. Markets achieve 71% accuracy forecasting which studies will replicate, beating surveys at 58% (Dreber et al., 2015). Top-market-price studies replicate 83% of the time; bottom-ranked hit 33% (Holzmeister et al., 2025).

Anomaly detection. AI systems that can't verify claims can still flag anomalies (claims deviating from established patterns) for prioritized human review. This inverts AI's usual role: instead of generating claims, it identifies which ones most warrant human attention.

The core insight: verification need not be all-or-nothing. A tiered system that scrutinizes high-stakes claims intensely while provisionally accepting low-stakes ones manages limited resources far better than uniform treatment.

3.2 Verification Cascades

We introduce *verification cascades*: hierarchical architectures that apply increasingly rigorous verification to claims passing initial filters. The concept borrows from software engineering. As Kleppmann (2025) observes, programmers trust compiler-verified code without examining machine output. Verification substitutes for understanding when the verification process itself is reliable.

The key insight is that AI can flag anomalies without reliably verifying claims, inverting its usual role: filtering what warrants human attention rather than replacing human judgment. A scientific verification cascade might proceed through four levels:

1. **Automated consistency** (syntax): logical coherence, citation validity, statistical checks
2. **AI-assisted plausibility** (semantic): alignment with established knowledge, anomaly flagging
3. **Expert review** (domain verification): technical validity assessment by specialists
4. **Experimental replication** (ground truth): empirical confirmation

Each level filters. Automated checks catch logical inconsistencies; AI assessment flags implausible claims; experts evaluate technical validity; replication confirms empirical claims. Most claims stop at lower levels; only high-stakes ones proceed to replication.

Worked example. An AI proposes that compound X inhibits protein Y, a potential antimicrobial target. *Level 1*: automated checks verify valid protein structure citations and consistent nomenclature. Pass. *Level 2*: AI cross-references Y's known binding sites and flags that X's molecular weight exceeds typical inhibitors for this class. Annotated "anomalous, requires expert review." *Level 3*: a medicinal chemist notes that while atypical, allosteric inhibition remains plausible. Pass with "provisional-expert-reviewed" status. *Level 4*: six months later, two independent labs report successful inhibition assays. The claim becomes "verified"; downstream claims building on it inherit elevated credibility.

Full verification is impossible for most claims. This architecture ensures consequential ones aren't accepted without scrutiny. The calibration challenge is real: too permissive and false claims propagate; too restrictive and valid claims get blocked.

3.3 Provisional Knowledge for AI-Generated Claims

Teller (2009) argued that scientific knowledge involves idealization and context-dependent accuracy standards, not absolute truth. Freiman (2023) proposes that AI-derived beliefs constitute a distinct epistemic category, "technology-based beliefs," neither instrument-based nor testimony-based. Building on these insights, we propose *provisional knowledge* as a formal category for AI-generated claims that are plausible but unverified. Neither verified knowledge nor mere speculation.

Provisional knowledge serves three functions:

Honest labeling. Claims carry explicit verification status, preventing inadvertent treatment as established fact. Metadata tracks: (a) generation source (human/AI/hybrid), (b) verification level achieved, (c) time since generation, (d) downstream dependencies.

Action under uncertainty. Researchers can build on provisional claims while tracking dependencies, so verification or falsification propagates to dependent work. Think version control for epistemic status.

Accountability through visibility. Claims staying provisional indefinitely get implicitly deprioritized; claims repeatedly cited get escalated for verification via epistemic triage.

Status transitions. The framework requires explicit conditions for status changes: (1) *Provisional* $\rightarrow$ *Verified*: independent replication by two or more labs, or formal proof verification for theoretical

claims; (2) *Provisional* → *Retracted*: failed replications exceeding a threshold (e.g., two independent failures), or demonstrated contradiction with verified claims; (3) *Provisional* → *Archived*: no citation, replication attempt, or downstream development after a set period (e.g., five years), meaning the claim lacks sufficient interest to warrant tracking. Thresholds are domain-dependent: drug-target interactions need stricter standards than exploratory computational predictions. But the framework demands explicit, pre-registered criteria, not ad hoc judgments.

This requires institutional infrastructure: citation practices that distinguish dependence on verified versus provisional claims, and norms separating building on provisional claims (acceptable with acknowledgment) from treating them as established (inappropriate without verification).

4 Related Work

Reproducibility Crisis. The replication crisis predates AI: Ioannidis (2005) argued most published findings are false; the Open Science Collaboration found only 36% of psychology studies replicate (Open Science Collaboration, 2015). AI introduces both new challenges and potential solutions.

Social Epistemology of Science. Hardwig (1991) established that scientific knowledge depends on testimony and trust. Shapin (1994) showed it is constituted through social credibility assessment. Koskinen (2024) argues we lack a satisfactory social epistemology for AI-based science.

AI for Scientific Discovery. Recent systems compress years of research into days (Gottweis et al., 2025; Lu et al., 2024), but evaluation reveals significant gaps: 42% experiment failure rates (Beel et al., 2025) and only 21% error detection recall (Son et al., 2025).

Epistemic Frameworks. Teller (2009) introduced provisional knowledge; Freiman (2023) proposes "technology-based beliefs" as a distinct epistemic category. Our contribution operationalizes these with explicit status transitions and triage mechanisms.

5 Risks and Limitations

The verification bottleneck poses risks beyond individual false claims. Our deeper concern is *epistemic pollution*: a state where the scientific corpus becomes so contaminated with unverified claims that reliable ones can no longer be distinguished (Singh, 2025).

Epistemic pollution operates through citation propagation (retracted papers keep receiving positive citations (Schneider et al., 2020)), training data contamination (future AI learns from valid and invalid claims alike), and epistemic learned helplessness (researchers abandoning verification entirely). Our frameworks are management strategies, not solutions; whether they prove adequate depends on implementation and how AI capabilities evolve.

6 Conclusion

Knowledge production and verification capacity are decoupling. We've proposed epistemic triage, verification cascades, and provisional knowledge as tools for navigating this bottleneck. The limit of autonomous reasoning isn't generation but verification: AI systems that can't reliably verify their own outputs require human oversight. These frameworks are themselves provisional starting points. The problem is not provisional. How science adapts will determine whether AI augments human knowledge or merely floods it.

References

Joeran Beel, Min-Yen Kan, and Moritz Baumgart. Evaluating Sakana's AI Scientist for autonomous research. *arXiv preprint arXiv:2502.14297*, 2025. doi: 10.48550/arXiv.2502.14297.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, et al. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, 2018. doi: 10.1038/s41562-017-0189-z.

Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015. doi: 10.1002/asi.23329.

Cristina Cornelio et al. The need for verification in AI-driven scientific discovery. *arXiv preprint arXiv:2509.01398*, 2025.

Anna Dreber, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347, 2015. doi: 10.1073/pnas.1516179112.

Ori Freiman. Analysis of beliefs acquired from a conversational AI: Instruments-based beliefs, testimony-based beliefs, and technology-based beliefs. *Episteme*, 21(3):1031–1047, 2023. doi: 10.1017/epi.2023.12.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025. doi: 10.48550/arXiv.2502.18864.

John Hardwig. The role of trust in knowledge. *The Journal of Philosophy*, 88(12):693–708, 1991. doi: 10.2307/2027007.

Felix Holzmeister, Magnus Johannesson, Robert Böhm, Anna Dreber, Jürgen Huber, and Michael Kirchler. Examining the replicability of online experiments selected by a decision market. *Nature Human Behaviour*, 9:316–330, 2025. doi: 10.1038/s41562-024-02062-9.

John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8):e124, 2005. doi: 10.1371/journal.pmed.0020124.

Matthew Kelly. The epistemic suite: A post-foundational diagnostic methodology for assessing AI knowledge claims. *arXiv preprint arXiv:2510.24721*, 2025.

Martin Kleppmann. Prediction: AI will make formal verification go mainstream. Blog post, martin.kleppmann.com, 2025.

Inkeri Koskinen. We have no satisfactory social epistemology of AI-based science. *Social Epistemology*, 38(4):458–475, 2024. doi: 10.1080/02691728.2023.2286253.

Michail Kovanis, Raphaël Porcher, Philippe Ravaud, and Ludovic Trinquart. The global burden of journal peer review in the biomedical literature. *PLOS ONE*, 11(11):e0166387, 2016. doi: 10.1371/journal.pone.0166387.

Keigo Kusumegi, Xinyu Yang, Paul Ginsparg, Mathijs de Vaan, Toby Stuart, and Yian Yin. Scientific production in the era of large language models. *Science*, 390(6779):1240–1243, 2025. doi: 10.1126/science.adw3000.

Jennifer Lackey. *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press, Oxford, 2008. doi: 10.1093/acprof:oso/9780199219162.001.0001.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. doi: 10.48550/arXiv.2408.06292.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349 (6251):aac4716, 2015. doi: 10.1126/science.aac4716.

Jodi Schneider, Di Ye, Alison M. Hill, and Ashley S. Whitehorn. Continued post-retraction citation of a fraudulent clinical trial report. *Scientometrics*, 125(3):2877–2913, 2020. doi: 10.1007/s11192-020-03631-1.

Marta Serra-Garcia and Uri Gneezy. Nonreplicable publications are cited more than replicable ones. *Science Advances*, 7(21):eabd1705, 2021. doi: 10.1126/sciadv.abd1705.

Steven Shapin. *A Social History of Truth: Civility and Science in Seventeenth-Century England*. University of Chicago Press, Chicago, 1994. doi: 10.7208/9780226148847.

Bhavneet Singh. Epistemic destabilization: AI-driven knowledge generation and the collapse of validation systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, volume 8, pp. 2387–2398, 2025.

Guijin Son, Jiwoo Hong, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha Choi, Gonçalo Paulo, Youngjae Yu, and Stella Biderman. When AI co-scientists fail: SPOT—a benchmark for automated verification of scientific research. *arXiv preprint arXiv:2505.11855*, 2025. doi: 10.48550/arXiv.2505.11855.

Paul Teller. Provisional knowledge. In Michel Bitbol, Pierre Kerszberg, and Jean Petitot (eds.), *Constituting Objectivity*, volume 74 of *The Western Ontario Series In Philosophy of Science*, pp. 503–514. Springer, Dordrecht, 2009. doi: 10.1007/978-1-4020-9510-8_30.

A Appendix

LLMs were used for drafting assistance during the preparation of this paper. The authors take full responsibility for all content.