CRITICAL NARRATIVE REVIEW / COMPUTER & COGNITIVE SCIENCES

Can Artificial Intelligence Understand What It Does Not Know?

Metacognition, Scientific Reasoning, and the Limits of Machine Intelligence

Milo Dolci[1]*

[1] Prisca Sapientia Institute, Zurich, Switzerland | director@priscasapientia.university

CORRESPONDING AUTHOR

Milo Dolci · director@priscasapientia.university

PUBLICATION

Volume 2 · Issue 1 · 2026

ABSTRACT

Background: Large language models (LLMs) can generate fluent explanations, estimate confidence, and critique their own outputs, yet the relation between these behaviors and genuine metacognitive competence remains unresolved. Objective: To critically examine whether artificial intelligence can identify the boundaries of its knowledge and how this capability should be evaluated in scientific settings. Approach: This critical narrative review synthesizes selected foundational and empirical research on calibration, verbalized uncertainty, hallucination, self-evaluation, self-correction, and evidence-grounded reasoning. It is not a systematic review or an original experimental study. Findings: Model confidence can be informative under particular evaluation protocols, but apparent certainty is not a dependable proxy for correctness. Performance depends on task framing, domain shift, feedback quality, and the availability of external evidence. Intrinsic self-correction often fails, whereas verification using reliable tools or external information can improve outcomes. Interpretation: Operational metacognition should be defined by measurable error detection, calibration, selective prediction, and appropriate abstention rather than by anthropomorphic descriptions of machine awareness. Conclusion: Scientific use of LLMs requires explicit uncertainty management, traceable evidence, reproducible validation, and accountable human oversight.

KEYWORDS

Artificial intelligence; large language models; metacognition; uncertainty calibration; epistemic humility; hallucination; self-correction; scientific reasoning; selective prediction.

© 2026 World Health Academy Publishing House

1. INTRODUCTION

Large language models have altered the practical meaning of asking a machine a question. A system can summarize a paper, construct a plausible hypothesis, produce code, and defend an explanation in polished prose. Yet none of these performances, by itself, establishes that the system can reliably distinguish what it knows from what it merely appears to know. The gap between linguistic competence and epistemic reliability is especially consequential in research, medicine, law, and education.

The question posed in this review is deliberately operational: can an AI system identify, communicate, and act on uncertainty about the correctness of its own outputs? This differs from asking whether machines possess consciousness or subjective experience. Those philosophical questions cannot be resolved by a model describing itself, and the present article makes no claim about machine phenomenology.

An adequate answer must distinguish at least four capacities: producing an answer, estimating its probability of being correct, identifying evidence that would settle the question, and changing behavior when the evidence is insufficient. A system may perform well on one dimension and poorly on another.

2. SCOPE AND REVIEW APPROACH

This article is a critical narrative synthesis of selected peer-reviewed papers and influential preprints on neural-network calibration, language-model uncertainty, hallucinations, self-correction, and evaluation. The selection emphasizes conceptual relevance, clearly reported experimental findings, and methodological contrasts rather than comprehensive database coverage. No systematic search protocol, risk-of-bias scoring, meta-analysis, or pooled effect estimate was conducted.

We use the term metacognitive performance to denote observable monitoring and regulation of model outputs. We reserve metacognitive experience for subjective phenomena, which cannot be inferred from the experiments reviewed here. This distinction prevents a measurable engineering property from being mistaken for a settled theory of mind.

3. WHAT DOES IT MEAN FOR A MODEL TO KNOW?

In human epistemology, knowledge is often discussed in relation to truth, justification, and the possibility of error. For machine learning, the more tractable question is whether an internal or reported score tracks the frequency with which predictions are correct. This property is called calibration. If predictions assigned a confidence of 80% are correct approximately 80% of the time within a defined evaluation set, the predictions are calibrated for that setting.

Calibration is not the same as accuracy: a model can be accurate but overconfident, or calibrated yet insufficiently accurate for clinical use. Guo and colleagues demonstrated substantial miscalibration in modern neural networks and examined temperature scaling as a practical corrective technique [1]. The underlying lesson extends beyond classification: confidence must be evaluated against outcomes, not inferred from eloquence.

For generative models, the situation is harder because answers are open-ended and correctness may be ambiguous, partially true, or dependent on context. A confidence score attached to a sentence may refer to lexical probability, answer correctness, source support, or the model’s propensity to produce that answer. These are distinct quantities.

4. SELF-EVALUATION AND VERBALIZED UNCERTAINTY

Kadavath and colleagues studied whether language models could estimate the validity of their own answers, including probabilities of a proposed answer being true and of knowing an answer before generation. Their experiments reported promising self-evaluation under specific formats, alongside limitations when transferring calibration across tasks [2]. Because the study is a preprint, its findings should be weighed accordingly.

Lin and colleagues found that language models could be trained to express meaningful uncertainty in words, with verbalized confidence exhibiting calibration on evaluated tasks [3]. This is significant because users encounter textual statements such as “I am moderately confident” rather than internal logits. However, a trained linguistic confidence report should not be mistaken for an introspective account of subjective certainty.

The crucial evaluation question is behavioral: does an uncertainty statement predict correctness on unfamiliar, consequential questions? A convincing test requires held-out topics, independently verified ground truth, repeated sampling, and measures of calibration and discrimination. It also requires recording whether a model declines to answer when confidence falls below a threshold.

5. HALLUCINATION AND THE ILLUSION OF CERTAINTY

Hallucination refers broadly to generated content that is unsupported, inconsistent with source material, or factually incorrect. Ji and colleagues catalogued the problem across natural-language generation tasks and discussed its evaluation and mitigation [4]. In scholarly writing, hallucinations may appear as invented references, fabricated quotations, distorted numerical results, or unwarranted causal conclusions.

Fluency creates a particular hazard: readers often use coherent language as a heuristic for expertise. An LLM may therefore sound more certain than its evidential position warrants. The risk is not limited to spectacular falsehoods; small distortions in study populations, endpoints, or confidence intervals can change the interpretation of an otherwise plausible literature review.

Uncertainty-aware generation should not simply add cautious adjectives to every sentence. It should separate verified statements, contested findings, hypotheses, and unknowns; provide accessible supporting evidence; and abstain when a claim cannot be responsibly substantiated.

6. CAN AI CORRECT ITS OWN MISTAKES?

Self-correction is attractive because it promises greater reliability without additional human work. However, an instruction such as “check your answer” may generate a second fluent response without introducing new evidence. Huang and colleagues found that intrinsic self-correction could fail and sometimes reduce reasoning performance in their experimental settings [5].

A critical survey by Kamoi and colleagues distinguished self-correction with internal feedback from approaches using external evidence or reliable task-specific verification. It concluded that success depends strongly on the feedback regime and evaluation design [6]. This is not a universal impossibility result: Wu and colleagues described improvements from structured key-condition verification on selected reasoning tasks [7].

These findings are better understood as a design principle than as a contradiction. Repeated generation is not equivalent to independent verification. A correction process is stronger when it can compare a claim against an external database, executable test, primary paper, or other trustworthy reference.

7. SCIENTIFIC REASONING IS MORE THAN PLAUSIBLE EXPLANATION

Scientific reasoning requires more than producing hypotheses that resemble scientific discourse. It involves identifying testable predictions, choosing discriminating experiments, tracking provenance, comparing alternative explanations, and updating conclusions in response to results. A model can assist with each activity, but the scientific validity of its output depends on the quality of the underlying evidence and method.

An LLM may propose a mechanistic explanation for a clinical association, but the explanation is not established merely because it is biologically plausible. Similarly, a generated statistical interpretation is not valid unless the design, data, assumptions, and computations are checked. In high-stakes domains, the model should distinguish the epistemic status of a claim from its rhetorical attractiveness.

One practical safeguard is to require an evidence ledger: every important claim is linked to a source, a source type, an evidence strength assessment, and a verification status. Unsupported claims remain visibly provisional. This procedure is a proposed editorial control, not a validated guarantee of factual accuracy.

8. AN AI SYSTEM EXAMINING THE LIMITS OF ITS OWN REASONING

An AI-generated analysis of AI uncertainty offers a useful methodological illustration but not privileged access to the model’s internal state. A language model can explain common failure modes, compare published findings, and articulate conditions under which it should defer to evidence. It cannot establish the truth of its own explanation solely by producing it.

The distinction can be made explicit through a simple thought experiment. Suppose a model answers a specialized question and then assigns itself 95% confidence. Without independently known ground truth, the confidence statement is just another model output. If the answer is checked against a reliable primary source, the system’s assessment becomes testable. Repeating this process over many questions allows estimation of calibration and error detection.

This section is therefore a conceptual self-audit of a class of systems, not an empirical experiment conducted on the model generating the manuscript. No claim of subjective awareness, private introspection, or experimentally demonstrated self-knowledge follows from it.

9. A PRACTICAL FRAMEWORK FOR EPISTEMIC RELIABILITY

We propose five linked requirements for trustworthy research assistance. First, claim-level provenance: consequential statements should be traceable to verifiable sources. Second, calibrated uncertainty: expressed confidence should correspond to observed correctness on representative tasks. Third, selective prediction: systems should be able to defer, abstain, or request clarification when evidence is inadequate.

Fourth, independent verification: calculations, citations, and factual assertions should be checked with tools or human review rather than by repeated generation alone. Fifth, post-deployment monitoring: performance should be evaluated as data, user populations, and model versions change. These requirements are complementary; no single confidence metric establishes reliability.

For evaluation, we recommend reporting accuracy or task-specific validity alongside expected calibration error, Brier score where applicable, risk–coverage curves for abstention, and subgroup analyses. Any such metrics must be interpreted within the constraints of the task and ground-truth construction.

10. IMPLICATIONS FOR SCIENTIFIC PUBLISHING

Generative AI can accelerate literature triage, language editing, code drafting, and manuscript structuring. These benefits do not remove the need for accountable human authorship and independent verification. Scientific editors should be able to distinguish author conclusions from generated suggestions and identify whether bibliographic details were checked against publisher records.

An editorial workflow should document which tasks involved AI, whether external sources were consulted, and who verified the final statements. Human authors must retain responsibility for originality, accurate citation, ethical compliance, and the integrity of the published record. An AI tool should not be listed as a human author merely because it produced text.

For the present manuscript, the declared AI contribution is substantive drafting and organization. Human review and approval are still required before submission; no automated drafting system can independently certify that the article is publication-ready.

11. LIMITATIONS

This narrative review is selective and does not quantify the overall prevalence of calibration failures or the magnitude of benefits from any intervention. The literature evolves rapidly, and results obtained with particular model versions, prompts, datasets, and evaluation protocols may not generalize to newer systems. Several influential studies concern benchmarks rather than real-world scientific workflows.

The article also cannot determine whether machine consciousness is possible. Its claims concern observable prediction, monitoring, and correction. The proposed reliability framework is normative and should be tested empirically before it is treated as a validated standard.

12. CONCLUSIONS

Artificial intelligence can display useful forms of uncertainty estimation and error monitoring, but these capabilities are conditional, measurable, and fallible. A fluent explanation is not evidence that a system knows its answer is correct. Likewise, a self-critique is not an independent audit when it relies on the same unverified information as the original response.

The scientifically productive question is therefore not whether a model “feels uncertain,” but whether it can reliably detect error, express uncertainty in calibrated terms, seek decisive evidence, and decline unsupported claims. Future progress should be judged by reproducible improvements in these behaviors and by the quality of human–AI research decisions—not by increasingly persuasive descriptions of machine self-awareness.

REFERENCES

1. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. Proceedings of ICML. 2017;70:1321–1330. https://proceedings.mlr.press/v70/guo17a.html

2. Kadavath S, et al. Language Models (Mostly) Know What They Know. arXiv preprint. 2022; arXiv:2207.05221. https://arxiv.org/abs/2207.05221

3. Lin S, Hilton J, Evans O. Teaching Models to Express Their Uncertainty in Words. Transactions on Machine Learning Research. 2022. https://mlanthology.org/tmlr/2022/lin2022tmlr-teaching/

4. Ji Z, Lee N, Frieske R, et al. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys. 2023;55(12):248. doi:10.1145/3571730

5. Huang J, Chen X, Mishra S, et al. Large Language Models Cannot Self-Correct Reasoning Yet. International Conference on Learning Representations. 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html

6. Kamoi R, Zhang Y, Zhang N, Han J, Zhang R. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. Transactions of the Association for Computational Linguistics. 2024;12:1417–1440. doi:10.1162/tacl_a_00713

7. Wu Z, Zeng Q, Zhang Z, Tan Z, Shen C, Jiang M. Large Language Models Can Self-Correct with Key Condition Verification. Proceedings of EMNLP. 2024:12846–12867. doi:10.18653/v1/2024.emnlp-main.714

DECLARATIONS

Article type: Critical narrative review; no original participant data or experimental results.

Author and corresponding author: Milo Dolci, Prisca Sapientia Institute, Zurich, Switzerland. Final author approval required.

AI-use disclosure: A generative AI system (OpenAI ChatGPT) assisted with topic formulation, synthesis, drafting, and layout. A qualified human author must verify every factual statement and reference, revise the manuscript, and accept accountability before submission.

Funding, conflicts of interest, and author contributions: To be supplied by the responsible human authors.

Editorial status: Draft for scholarly assessment; not evidence of journal acceptance or publication.