Who Gets to Contest Science? Generative AI and the Democratization of Scientific Scrutiny

George Kujo

Independent Researcher | Tokyo, Japan

ORCID: 0009-0006-6864-0145
george.kujo.irjp@gmail.com

Preprint | v1.0 | 6 September 2026

Abstract

Generative AI is increasingly discussed as a technology that may democratize scientific production by lowering barriers to writing, coding, and analysis. Less attention has been paid to a parallel possibility: that it may lower barriers to scrutinizing scientific claims outside established evaluative roles. This article develops a conceptual account of that shift. It distinguishes access to scientific materials from the capability to interrogate them and argues that generative AI may contribute to a dual democratization of scientific production and scientific scrutiny. By assisting with statistical interpretation, code inspection, literature comparison, citation checking, and methodological reasoning, AI may distribute some evaluative capabilities beyond formal peer reviewers and institutionally recognized experts. This creates a capability–recognition mismatch: who can identify a potentially valid scientific problem may diverge from whose criticism institutions recognize. Yet expanded scrutiny also creates a quality-control problem. The same systems can lower the cost of pseudo-scrutiny, hallucinated objections, and technically sophisticated but unverifiable claims. The institutional challenge is therefore how to distinguish objections that merit escalation from those that do not. I argue for a layered governance principle: relatively open participation in detecting scientific problems, combined with progressively stronger requirements for provenance, reproducibility, evidentiary specificity, and expert adjudication before institutional consequences follow. Generative AI may democratize science not only by changing who can produce scientific claims, but also by changing who can practically challenge them.

Keywords: Generative AI; Scientific scrutiny; Open Science; Evaluative capability; Institutional recognition; Scientific governance

1. Introduction

Generative artificial intelligence is rapidly becoming embedded in scientific work. Large language models (LLMs) and related systems can assist with literature retrieval and synthesis, coding, data analysis, manuscript preparation, hypothesis generation, and increasingly complex components of scientific discovery (Reddy & Shojaee, 2025; Zhang et al., 2025). Recent research suggests that these tools are not merely changing how individual tasks are performed but are beginning to alter patterns of scientific production itself. In a study of AI tools across distinct technological eras, Hao et al. (2026), analyzing more than 41 million scientific papers, found that the adoption of AI tools was associated with substantially greater publication and citation output among individual scientists, even while scientific activity became more concentrated around established areas of inquiry. More broadly, generative AI has been described as a potential democratizing technology for knowledge work because it can make cognitively demanding tasks accessible to users who previously lacked some of the specialized skills required to perform them (Daepp et al., 2026). Much of the discussion surrounding AI and science has accordingly focused on an increasingly familiar question: *who will be able to produce scientific knowledge?*

Yet scientific systems depend not only on the production of claims but also on their criticism. Published findings must be interpreted, checked against prior literature, examined for analytical weaknesses, reproduced where possible, challenged when warranted, and corrected when they fail. Existing discussions of generative AI have extensively considered research productivity, scientific discovery, manuscript preparation, and, more recently, the use of AI inside peer review. A 2026 scoping review of 189 sources on AI in scholarly peer review, for example, documents applications ranging from editorial triage and reviewer assistance to automated review generation (Nabavi et al., 2026). These developments are consequential, but they predominantly concern the use of AI *within an already established evaluative role*. Less attention has been paid to a different possibility: that generative AI may alter who possesses the practical capability to scrutinize scientific claims outside those formal roles.

This possibility cannot be reduced to Open Science or open participation. Open Science has already sought to make scientific publications, data, software, infrastructures, and research processes more accessible and reusable, while citizen science and other participatory models have widened opportunities for societal actors to participate in knowledge production (UNESCO, 2021). Open peer review has likewise included forms of "open participation" in which members of a wider community can contribute evaluative comments rather than leaving criticism exclusively to invited reviewers (Ross-Hellauer, 2017). Post-publication platforms such as PubPeer further demonstrate that scrutiny need not end when a manuscript is accepted (PubPeer, n.d.). Nevertheless, an opportunity to participate is not the same as a practical capability to participate. A freely accessible regression table remains difficult to interrogate without statistical knowledge; open source code remains difficult to evaluate without programming competence; an available supplementary appendix may remain effectively inaccessible to a reader who cannot reconstruct its analytical logic. Citizen-science verification practices likewise show that broad participation can coexist with checks on the quality of distributed contributions (Baker et al., 2021). *Formal openness therefore does not necessarily produce effective evaluative participation.*

Generative AI potentially intervenes at this second layer. It can translate technical language, explain statistical procedures, generate and interpret code, summarize competing literatures, identify claims and citations, suggest robustness checks, and assist users in reconstructing analytical procedures. Experimental evidence already shows that AI-generated simplification can improve lay readers' comprehension of scientific material (Markowitz, 2024), although comprehension alone is far from equivalent to scientific

expertise. This assistance may reduce the cognitive and technical barriers required to engage meaningfully with scientific claims. The distinction is important: open access changes whether a person can reach an article; AI assistance may change what that person can practically do after reaching it. In principle, the same technological shift that lowers barriers to producing a scientific claim may therefore lower some of the barriers to interrogating one.

I call the latter possibility *AI-assisted distributed scrutiny*. It should be distinguished from AI-assisted peer review. AI-assisted peer review concerns the use of AI by actors already occupying an established evaluative position—for example, an editor using AI for triage or a reviewer using an LLM to inspect a manuscript. AI-assisted distributed scrutiny concerns the use of AI by a wider set of human actors—researchers outside the immediate field, practitioners, independent researchers, journalists, patient or public-interest groups, citizen scientists, and technically capable members of the public—to understand, verify, reproduce, challenge, or otherwise interrogate published scientific claims. The novelty is not that outsiders can criticize science; they have long done so. Nor is it that post-publication review exists outside conventional peer review. The relevant change is that generative AI may reduce some of the capability barriers that previously limited how deeply and at what scale such actors could engage with technical scientific material.

If evaluative capability becomes more widely distributed, however, scientific authority does not automatically become equally distributed. Scientific criticism has traditionally acquired credibility partly through institutional roles and status signals: disciplinary expertise, academic credentials, institutional affiliation, editorial selection, peer recognition, and established track records. These signals are imperfect, but they help solve a longstanding recognition problem in social epistemology: people who cannot independently evaluate expertise must nevertheless decide whose scientific judgments deserve weight. Recent work continues to show both the importance and the limitations of credentials and related indicators when non-experts assess epistemic trustworthiness (Mirkin & Sojka, 2026). Generative AI may therefore create an asymmetry. The practical capacity to interrogate scientific claims may spread beyond the institutional roles through which legitimate scientific criticism has traditionally been recognized. I describe this not as the disappearance of expertise, but as a possible *capability–recognition mismatch*.

Such a development would not imply that more criticism necessarily produces more reliable science. AI can lower the cost of valid scrutiny, but it can also lower the cost of superficial objections, fabricated citation challenges, methodological cargo cults, and highly confident pseudo-scrutiny. Evidence from lay use of LLMs already suggests that access to an AI system does not by itself guarantee evidence-grounded evaluation: users can obtain poor-quality outputs, and even simple interventions that improve their use of these systems do not eliminate substantial quality deficits (Rebitschek et al., 2025). Post-publication review itself has also created substantial institutional processing burdens (Caron et al., 2025). The central governance problem is therefore not how to recognize every challenge as equally valid, but how to distinguish *credible scrutiny* from merely inexpensive scrutiny. This article develops that problem through three questions: how generative AI changes the capability barriers to producing and scrutinizing scientific claims; how AI-assisted distributed scrutiny differs from Open Science, citizen science, conventional peer review, and post-publication review; and what follows when evaluative capability becomes more widely distributed than the institutional recognition traditionally attached to scientific evaluation. I argue that generative AI may democratize science in a second and insufficiently examined sense—not only by widening who can participate in producing scientific claims, but by widening who can meaningfully challenge them. *The resulting question is not merely who can produce science, but who gets to contest it.*

The argument comprises five related propositions, whose numbering marks their conceptual roles. P1 states the dual-democratization thesis: AI lowers capability barriers to scientific production and scrutiny. P2 distinguishes access to materials from the capability to examine them. P3 specifies the scrutiny-side mechanism of P1, identifying how lower task costs can extend substantive evaluation beyond assigned evaluative roles; it is a corollary of that thesis, not an independent democratization claim. P4 identifies the possible gap between this capability and institutional recognition. P5 identifies the evidentiary requirements for credible distributed scrutiny. The exposition establishes the access–capability distinction first (P2, Section 2), then develops the overarching thesis (P1, Section 3), its scrutiny-side corollary (P3, Section 4), and its institutional implications (P4 and P5, Sections 5 and 6).

2. From Access to Capability

The democratization of science has historically been pursued primarily by expanding access and participation. Open-access publishing reduces price barriers to scientific literature. Open-data and open-code initiatives make research materials available for inspection and reuse. Open Science more broadly seeks to increase the accessibility, transparency, and reusability of scientific processes and outputs (UNESCO, 2021; Wilkinson et al., 2016). Citizen-science initiatives extend participation beyond professional researchers by involving members of the public in activities ranging from data collection to interpretation and, in some cases, research design (UNESCO, 2021). Open peer review and post-publication review similarly widen the range of actors who may participate in scientific evaluation (Ross-Hellauer, 2017).

These developments represent substantial changes in the institutional accessibility of science. They do not, however, eliminate another class of barriers: those arising from the knowledge, skills, time, and technical competence required to make effective use of that access. A research article may be freely available but remain difficult to interpret. A dataset may be downloadable but practically unusable to someone unfamiliar with its structure or the statistical software required to analyze it. Source code may be publicly posted while remaining opaque to readers unable to understand the programming language, computational environment, or analytical assumptions embedded within it. Opening scientific objects therefore increases the opportunity to inspect them without necessarily distributing the capability required to do so.

This distinction is particularly important because participation has already been incorporated into existing models of scientific openness. Ross-Hellauer's (2017) systematic review of open peer review identified “open participation”—the ability of a wider community to contribute to review—as one of the recurring characteristics of open-review systems. The possibility that people outside a conventionally invited reviewer pool can comment on scientific work is therefore not novel to generative AI. Nor are public criticism, citizen participation, or post-publication commentary new phenomena. The relevant question is instead whether technological change alters the *depth of participation that is practically achievable* by people who already possess formal access to the material.

I therefore distinguish *access* from *capability*. Expanding access concerns whether scientific information, materials, or evaluative processes are available to a wider population. Lowering capability barriers concerns whether the cognitive and technical requirements for performing meaningful scientific work with those materials become less demanding. The distinction is analytical rather than absolute. Access is often a precondition for capability, and capability can itself depend on education, disciplinary experience, institutional support, software, data infrastructure, and time. Yet the two are not equivalent. A system can be highly open while remaining practically usable only by a narrow group possessing the skills necessary to exploit that openness.

Generative AI is relevant because it can intervene between formal access and practical use. A reader confronted with an unfamiliar statistical method can request an explanation of the estimator, its assumptions, and common failure modes. Code written in an unfamiliar language can be translated, annotated, debugged, or reconstructed. A supplementary appendix can be summarized and connected to the claims made in the main article. Citations can be extracted for checking. Alternative specifications can be proposed. Technical terminology can be translated into more accessible language. These functions do not eliminate the need for judgment, domain knowledge, or verification, but they can reduce the cost of acquiring the intermediate understanding necessary to begin an evaluation.

Evidence from science communication illustrates the narrower version of this mechanism. Markowitz (2024) found that GPT-generated simplifications of scientific material could make scientific information easier for lay readers to process and altered their perceptions of scientists. This finding should not be interpreted as evidence that a language model converts a lay reader into an expert. It demonstrates something more limited but theoretically important: at least some cognitive barriers separating technical scientific information from non-specialist readers are technologically malleable. The result concerns comprehension rather than research evaluation, but it shows that the practical intelligibility of technical material is not fixed.

This matters for scientific scrutiny because evaluation rarely consists of a single act of expert judgment. It is composed of many smaller operations: identifying a paper's central claim, reconstructing how an analysis leads to that claim, locating the relevant comparison literature, checking whether a cited source supports the proposition attributed to it, understanding a statistical model, running provided code, changing an analytical assumption, or determining whether an apparent inconsistency warrants deeper investigation. Traditionally, many of these operations required direct possession of specialized technical skills or access to collaborators who possessed them. Generative AI can increasingly mediate some of these intermediate steps.

The resulting change is therefore better understood as a modification of the *capability threshold for scrutiny* than as the elimination of expertise. Consider three stylized cases. A journalist who previously could read the abstract of an epidemiological paper but not evaluate its regression specification may now obtain a structured explanation of the model and identify questions requiring expert follow-up. A clinician outside academic research may be able to inspect provided analytical code, reproduce an output, and test a simple alternative specification without having previously learned the relevant programming language in depth. An independent researcher may be able to trace dozens of cited claims, compare supplementary materials across studies, and reconstruct an analysis at a scale that would previously have required substantially more time or specialized assistance. In each case, AI does not confer disciplinary authority. It changes the cost structure of the tasks that precede and support evaluation.

This distinction also clarifies why *AI-assisted distributed scrutiny* should not be treated as another label for citizen science. Citizen science usually concerns participation in the production of scientific knowledge, although its forms vary substantially and may include interpretation or co-design. Distributed scrutiny concerns the evaluation of claims that have already entered, or are attempting to enter, the scientific record. Nor is it synonymous with open peer review. Open peer review changes who may participate, what is visible, or how the review process is organized. Distributed scrutiny concerns what those participants may be technically capable of doing once they engage. One describes an institutional opportunity structure; the other describes a changing capability structure.

The distinction can be represented as a simple two-dimensional model. *Access without capability* describes science that is formally open but practically difficult for most people to interrogate. *Capability without access* describes cases in which people possess relevant expertise but cannot inspect closed articles, unavailable data, proprietary code, or inaccessible supplementary material. *Access with capability* creates the strongest conditions for distributed scrutiny because both the scientific object and the practical means of interrogating it are available.

Generative AI does not solve the access problem, nor does it independently create the final condition. Its distinctive relevance is that it may shift individuals from the first condition toward the third by reducing some of the cognitive and technical costs that remain after formal openness has been achieved. This framework establishes P2, the access–capability distinction introduced above:

P2 — Access–Capability Distinction. Open access to scientific information is distinct from the practical capability to interpret, verify, and contest that information.

This distinction supplies a premise for the scrutiny-side corollary developed as P3 in Section 4. If generative AI can lower the capability threshold associated with tasks such as technical interpretation, code reconstruction, literature comparison, citation verification, and exploratory reanalysis, scientific evaluative capability may become distributed among a wider population than was previously able to exercise it at comparable cost.

This does not establish that the resulting evaluations will be correct. Lowering a capability threshold may increase both competent and incompetent uses of a tool. Generative AI can explain a statistical model correctly or incorrectly; reconstruct an analytical pipeline faithfully or introduce subtle errors; identify a genuine citation mismatch or hallucinate one. The distinction developed here therefore concerns *the capacity to undertake scrutiny, not the validity of its outcome*. That distinction will become central later in the paper: democratizing the ability to challenge scientific claims creates a new problem of evaluating the challenges themselves.

Open Science addresses access as a central bottleneck of scientific communication (UNESCO, 2021). Generative AI may now weaken a second one: the technical and cognitive threshold separating access to scientific material from the ability to interrogate it. If so, the consequences of AI for science cannot be understood only by counting how many more people can produce papers. They must also include how many more people can inspect, reproduce, question, and contest them.

3. Generative AI and Scientific Production

The most visible effect of generative AI on science has been its integration into research production. Large language models and related systems now support activities distributed across much of the research cycle, including literature synthesis, hypothesis generation, coding, experimental design, data analysis, interpretation, manuscript preparation, and scientific communication. Recent reviews of AI for scientific discovery describe a progression from task-specific assistance toward increasingly integrated systems capable of performing multiple stages of the scientific process (Reddy & Shojaee, 2025). Other work similarly characterizes large language models (LLMs) as potential productivity-enhancing components of the scientific method rather than as tools confined to writing or information retrieval (Zhang et al., 2025).

This expansion has stimulated a broader literature on scientific automation and autonomy. Contemporary systems have been proposed not only as assistants but as analysts and, in increasingly ambitious

architectures, as semi-autonomous or autonomous scientific agents. Reddy and Shojaee (2025) describe progress in scientific reasoning, simulation, and experimentation while identifying continuing obstacles to integrated, autonomous research. Zhang et al. (2025) likewise distinguish current assistance from the prospect of systems that contribute to fundamental discoveries. These developments leave open how far autonomous scientific judgment can be delegated to present systems.

The consequences are not limited to automation. Generative AI can reduce the amount of specialized labor required to perform tasks that historically demanded substantial training or collaboration. Programming assistance may allow researchers with limited software expertise to construct or modify analytical workflows. Language models can accelerate literature synthesis and help translate between disciplinary vocabularies. Tools for hypothesis generation can produce large numbers of candidate ideas, while AI-assisted analysis can reduce the time required to explore data or test alternative specifications. These capabilities have contributed to a growing discussion of AI as a democratizing technology for knowledge work. Daepp et al. (2026), for example, explicitly examine whether generative AI broadens access to high-complexity knowledge-work tasks and who is positioned to benefit from those capabilities.

The democratization claim nevertheless requires qualification. Lower task costs do not eliminate unequal access to data, infrastructure, compute, disciplinary knowledge, mentorship, research funding, or institutional networks. Daepp et al. (2026) identify social and place-based divides in the benefits of AI-assisted knowledge work. Nor does technical assistance ensure high-quality scientific reasoning: evaluation, reproducibility, and human validation remain constraints on AI-assisted discovery (Reddy & Shojaee, 2025).

For the present argument, however, it is unnecessary to resolve whether AI will become an autonomous scientist. The more conservative proposition is already sufficient: generative AI can reduce the cognitive, technical, and temporal costs associated with multiple components of scientific production. Whether those components are ultimately coordinated by a human researcher or an autonomous system, the threshold for carrying out many research tasks is changing. This observation motivates P1, the overarching dual-democratization thesis:

P1 — Dual Democratization. Generative AI can lower capability barriers to both scientific claim production and scientific scrutiny.

The production half of this proposition is already comparatively well developed in the literature. The scrutiny half is not. Much of the existing work asks whether AI can generate hypotheses, conduct analyses, write papers, or assist reviewers. These are important questions, but they leave aside a structurally different consequence of the same technological capabilities. The very tools that help a researcher construct an analysis can also help another person reconstruct it. A model that explains code to its author can explain the same code to a critic. A system that accelerates literature synthesis for manuscript preparation can also accelerate citation checking after publication. A tool that proposes an analytical specification can also suggest an alternative specification with which to challenge a published result.

The same reduction in task cost can therefore operate in two directions. *Production:* from question or dataset to scientific claim. *Scrutiny:* from scientific claim back to assumptions, evidence, code, citations, and alternative explanations. Existing discussions of AI and science have concentrated primarily on the first direction. The remainder of this paper concentrates on the second.

4. Generative AI and Scientific Scrutiny

4.1. Formal Peer Review as Institutionalized Scrutiny

Scientific criticism has traditionally been institutionalized most visibly through peer review. Before publication, editors select reviewers who are expected to assess the validity, significance, methodology, interpretation, and presentation of a manuscript. The reviewer therefore occupies two positions simultaneously: an evaluative position, because the reviewer is expected to scrutinize scientific claims, and an institutional position, because the journal has formally authorized that person to perform the scrutiny (Ross-Hellauer, 2017).

These two functions are analytically distinct. A reviewer does not acquire the technical ability to evaluate a manuscript merely by receiving an invitation. Rather, the journal uses disciplinary expertise, publication history, professional reputation, and other signals to identify people presumed to possess that ability. Peer review therefore links *evaluative capability* with *institutional recognition* through a selection mechanism.

That mechanism has important advantages. It concentrates evaluation among people expected to possess relevant expertise and provides editors with an administratively manageable set of judgments. Yet it also constrains the number and range of people who formally participate in pre-publication evaluation. Reviewer scarcity, workload, disciplinary specialization, conflicts of interest, and imperfect reviewer selection have long been recognized limitations of the system (Nabavi et al., 2026). A small number of reviewers cannot inspect every assumption, citation, analytical decision, dataset, or line of code in a modern research article. The publication decision consequently cannot be understood as the completion of scientific evaluation. It represents one institutional checkpoint in a longer process of scrutiny.

4.2. AI-Assisted Peer Review

Generative AI has increasingly entered this established evaluative system. Reviewers may use language models to summarize manuscripts, identify potential weaknesses, organize comments, improve wording, inspect statistical methods, or generate portions of review reports. Editors and publishers are also experimenting with AI-assisted triage, reviewer selection, technical checks, and automated evaluation.

The literature surrounding these practices has expanded rapidly. Nabavi et al. (2026) identified 189 sources concerning AI in scholarly peer review through August 2025, spanning editorial triage, reviewer assistance, and autonomous review generation. Their synthesis identifies efficiency benefits alongside risks involving confidentiality, bias, and inadequate domain reasoning. These concerns have generated an important governance literature. The review also describes fragmented publisher policies and limitations on autonomous evaluation (Nabavi et al., 2026). Fluent review-like text consequently does not itself establish the quality of the underlying judgment. For the argument developed here, however, AI-assisted peer review remains institutionally conservative in one important sense. It changes the tools available to the reviewer without necessarily changing who counts as a reviewer.

The basic structure remains:

journal → recognized reviewer → AI assistance → evaluative judgment.

The reviewer has already crossed the institutional threshold before AI enters the process. AI-assisted peer review concerns the use of AI within an established evaluative role. AI-assisted distributed scrutiny concerns the expansion of evaluative capability beyond that role.

4.3. Post-Publication Scrutiny

Scientific evaluation also occurs after publication. Letters to editors, commentaries, replication studies, corrections, social-media discussions, institutional investigations, dedicated post-publication review systems, and platforms such as PubPeer provide mechanisms through which published claims can be challenged after conventional peer review has ended (Hardwicke et al., 2022; PubPeer, n.d.).

Post-publication peer review is important precisely because pre-publication review is necessarily incomplete. Errors may become visible only after other researchers attempt replication, inspect underlying images or data, apply methods in different settings, or compare a paper with a broader literature. Online platforms further allow criticism to occur more rapidly and independently of the journal that originally published the work. PubPeer illustrates both the power and the institutional consequences of distributed post-publication scrutiny. Caron et al. (2025) describe how comments have led to research-misconduct proceedings and how their volume and velocity can burden institutional responses. Post-publication review therefore already demonstrates that scientific scrutiny can be distributed beyond the small group originally selected by a journal.

Yet distribution remains uneven. Participation in PubPeer or similar systems does not imply that every participant possesses an equal ability to reconstruct an analysis, examine a statistical model, inspect code, trace citations, or identify technically meaningful inconsistencies. Open participation expands the set of potential critics, but it does not remove the capability barriers described in Section 2. Generative AI potentially changes this relationship.

4.4. AI-Assisted Distributed Scrutiny

I use *AI-assisted distributed scrutiny* to describe scientific evaluation in which generative AI reduces some of the technical or cognitive costs faced by human actors scrutinizing scientific claims outside a formally assigned evaluative role.

Its defining feature is not the use of AI itself. AI-assisted manuscript review is already widespread enough to constitute a separate literature. Nor is its defining feature public participation, which predates generative AI through open peer review, citizen science, letters, blogs, replication communities, and post-publication review. The distinctive feature is the interaction of three conditions:

1. *the scientific claim is available for external examination;*

2. *the evaluator is not dependent on appointment to a formal reviewing role; and*

3. *AI mediates tasks that previously required greater technical expertise, time, or specialist assistance.*

Under these conditions, scrutiny becomes potentially more distributed because the practical cost of performing parts of the evaluative process falls.

This mechanism is already visible in narrower technical applications. Alvarez et al. (2024) demonstrated scientific claim verification using LLMs and examples derived from citation text, obtaining performance close to earlier fine-tuned systems trained on manually curated claim–evidence pairs. Edelman and Skolnick (2025) developed Valsci, an open-source system for large-batch scientific claim verification. In their evaluation, it reduced misclassification relative to base-model outputs, avoided fabricated citations, and operated substantially faster than manual review by an undergraduate researcher. These systems are usually

discussed as tools for researchers, publishers, reviewers, or research-integrity professionals. But the underlying functions are not intrinsically tied to those institutional roles.

A citation-verification tool does not need to know whether its user has been appointed as a reviewer. A code interpreter does not require an academic affiliation before explaining an analytical pipeline. A language model does not require editorial authorization before comparing a claim with its cited source. This separation between *evaluative function* and *evaluative role* is central to AI-assisted distributed scrutiny.

Consider citation checking. A paper may contain dozens or hundreds of references. Determining whether each reference actually supports the proposition attributed to it is conceptually straightforward but labor intensive. Automated extraction, retrieval, summarization, and claim–evidence comparison can reduce that cost dramatically. A task that might previously have been undertaken only selectively by a reviewer or highly motivated critic becomes more feasible at larger scale.

The same logic applies to analytical code. Publicly available code historically benefited readers already able to understand the relevant programming language and computational environment. A generative model can now explain functions line by line, translate between languages, identify dependencies, reconstruct missing steps, or generate modified specifications. This does not guarantee correct interpretation, but it decreases the amount of prior programming expertise required to begin interrogating the analysis.

Statistical interpretation provides another example. A reader who recognizes that a paper's conclusion depends on an unfamiliar estimator may previously have needed a statistician before proceeding further. An AI system can provide an initial explanation of assumptions, identify common threats to validity, suggest diagnostic questions, and generate code for sensitivity analyses. The reader may still require expert confirmation, but the point at which expert assistance becomes necessary can move further downstream.

Literature comparison can change similarly. An individual critic historically faced substantial costs when determining whether a paper accurately represented dozens of preceding studies. AI-assisted search, extraction, summarization, and comparison can compress that workload. The same infrastructure that allows authors to synthesize a literature faster can therefore allow critics to interrogate that synthesis faster.

These examples reveal an important asymmetry in much of the current discussion of AI for science. Scientific AI tools are commonly classified according to the task they perform—literature review, coding, statistical analysis, image inspection, citation checking, or hypothesis generation. But many of these tasks are *directionally neutral*. Code generation can assist production; code reconstruction can assist scrutiny. Literature synthesis can support an Introduction; literature comparison can test whether that Introduction is accurate. Statistical modeling can generate a result; model explanation and alternative specification can interrogate that result. Claim verification can support an author before submission; the same capability can support a critic after publication.

The technological capability does not inherently belong to either side of the scientific claim.

This gives distributed scrutiny its relationship to the dual-democratization model. Generative AI does not merely create new instruments for researchers. It makes some existing research instruments available at lower effective skill and time thresholds to anyone able to access them. The relevant population need not consist of people attempting to become substitute experts. Different actors may contribute different forms of scrutiny.

A clinician may notice that assumptions in a published model conflict with operational reality. A patient advocate may identify a discrepancy between a reported outcome and what the underlying measurement

actually captures. A journalist may trace a central claim through its citation chain. A researcher from another discipline may recognize a methodological assumption that is unusual outside the paper's immediate field. An independent researcher may reproduce an openly available analysis or compare published claims across sources. A technically skilled member of the public may detect an inconsistency that warrants expert examination. AI can increase the number of such actors capable of carrying an observation beyond intuition toward a structured, evidence-linked objection.

The transition can be represented as a ladder of scrutiny:

Observation "This result seems surprising." → *Interrogation* "What assumptions, data, and analyses produced it?" → *Reconstruction* "Can the reported pathway from evidence to conclusion be reproduced?" → *Challenge* "Does the claim survive an alternative specification, comparison, or source check?" → *Contestable criticism* "Can the objection be stated with evidence and procedures that other people can independently inspect?"

Generative AI can reduce costs at several transitions in this ladder. It is unlikely to eliminate the need for expertise at all of them. Its significance lies instead in allowing more actors to climb further than they previously could. This specifies P3 as a corollary of the scrutiny component of P1:

P3 — Distributed Scrutiny. Generative AI can expand the population capable of undertaking substantive components of scientific scrutiny beyond formal peer-review roles and traditional expert communities by lowering the technical, cognitive, and temporal costs of evaluative tasks.

The proposition deliberately concerns *components of scrutiny*, not equivalent expertise. That distinction prevents an important category error. Being able to run a sensitivity analysis does not establish competence to interpret all of its implications. Detecting an anomalous citation does not establish that the paper's conclusion is false. Producing a plausible methodological objection does not establish that the objection is valid. AI-assisted distributed scrutiny therefore redistributes the ability to generate *candidate criticisms* more readily than it redistributes the ability to establish their scientific validity. This difference becomes increasingly important as the cost of criticism falls.

If generating a methodological objection once required hours of specialist work but now requires minutes of AI-assisted investigation, the number of objections that can be produced may increase substantially. Some will identify real errors. Others will reflect misunderstanding, inappropriate model assumptions, fabricated evidence, or AI hallucination. The problem facing science may consequently change from a shortage of criticism to a shortage of mechanisms for distinguishing credible criticism from noise.

This is not entirely unprecedented. PubPeer already illustrates how increases in the volume and velocity of public scientific criticism can create institutional burdens. Generative AI potentially adds another scaling mechanism: not simply more people with opportunities to comment, but more people equipped with tools capable of producing technically elaborate challenges. The important consequence is therefore not that scientific authority becomes obsolete. It is that *evaluative capability can increasingly exist outside the institutional locations in which science has traditionally expected to find it*. That produces the next problem.

When a formally appointed reviewer raises an objection, the journal has already supplied a mechanism for receiving and weighting that objection. When an editor raises one, institutional authority is similarly clear. When an anonymous PubPeer commenter, independent researcher, journalist, clinician, patient advocate, or technically capable outsider produces a reproducible challenge to a published result, the mechanism for deciding how that criticism should count is much less settled. Generative AI may therefore solve part of one

problem while exposing another. It can lower the cost of asking: *"Can this claim be challenged?"* It does not answer: *"When should that challenge be recognized as scientifically credible?"*

The next section turns from distributed evaluative capability to this problem of scientific recognition.

5. Distributed Evaluative Capability and Scientific Recognition

Scientific scrutiny requires more than the capacity to formulate an objection. It also requires some mechanism through which objections can acquire epistemic weight. This creates a problem that becomes more visible once evaluative capability is separated from institutional role.

Scientific institutions have historically relied on status signals to help solve this recognition problem. Academic qualifications, publication records, disciplinary specialization, institutional affiliation, professional appointments, editorial selection, and peer recognition all provide imperfect but socially usable indicators of expertise. A journal editor cannot independently reconstruct the full competence of every potential reviewer. A clinician cannot personally verify the expertise of every scientist whose work informs clinical guidance. A member of the public cannot reproduce the entire body of research underlying every expert claim. Scientific communities therefore rely on social mechanisms for identifying people whose judgments are considered worthy of epistemic trust (Rolin, 2025).

These mechanisms are not arbitrary. Expertise is partly constituted through sustained disciplinary experience, tacit knowledge, methodological competence, and participation in specialist communities. Recent work on epistemic trust emphasizes precisely this point. Mirkin and Sojka (2026), for example, argue that commonly invoked markers such as credentials, track record, and dialectical performance do not allow non-experts to directly determine whether someone possesses expertise. They function instead as indirect indicators of experience and epistemic trustworthiness. Keren (2025) similarly distinguishes expertise from epistemic authority, emphasizing that authority concerning a particular question cannot be inferred from expertise alone. The institutional recognition of expertise therefore performs an important epistemic function. It reduces the cost of deciding whose judgments should initially receive attention.

The problem is that recognition and capability are not perfectly aligned. Recognized experts can be wrong. Experts can speak outside the domain in which their experience is pertinent. Credentialed researchers can misunderstand methods outside their own specialization, overlook errors, fail to disclose conflicts, or produce unreliable testimony. Recent social-epistemological work distinguishes pseudo-experts who lack genuine expertise from unreliable experts who possess legitimate credentials but nevertheless provide systematically unreliable testimony (Croce & Marsili, 2025). Credentials are therefore useful signals, but they do not guarantee the validity of a particular scientific claim or criticism. The reverse situation is also possible. A person may identify a valid and reproducible problem without occupying a role from which scientific institutions ordinarily expect authoritative criticism.

This possibility predates generative AI. Public and anonymous post-publication commenters have contributed concerns that subsequently entered formal research-integrity processes (Caron et al., 2025). PubPeer institutionalizes part of this possibility by allowing public and anonymous post-publication commentary. Importantly, PubPeer itself does not certify the scientific correctness of comments; readers are expected to evaluate them (PubPeer, n.d.). The platform therefore separates the opportunity to raise criticism from institutional certification of that criticism.

If the technical cost of reconstructing analyses, tracing citations, inspecting code, comparing literatures, or generating sensitivity checks falls, more people outside recognized evaluative roles may be able to produce objections that are technically substantive enough to warrant attention. The resulting challenge is not simply whether outsiders should be trusted. It is how scientific institutions should evaluate criticism when the traditional proxy—

recognized status → presumed evaluative competence

—becomes less tightly coupled to the practical capacity to perform specific evaluative tasks. I describe this as a *capability–recognition mismatch*: a condition in which the capacity to perform substantive components of scientific evaluation is distributed more broadly than the institutional recognition ordinarily used to identify legitimate evaluators. The concept does not imply that capability is equivalent to expertise. Nor does it imply that institutional recognition is obsolete. It identifies a change in the relationship between the two.

Before widespread generative AI, some forms of criticism were costly enough that the ability to produce them often correlated strongly with specialist training or institutional location. Reproducing a computational analysis might require substantial programming expertise. Checking a long citation chain might require days of manual retrieval. Exploring alternative statistical specifications might require access to a statistician. These costs functioned, imperfectly, as informal barriers separating technically elaborate criticism from ordinary commentary. As those costs fall, the signal carried by the mere existence of a technically elaborate criticism becomes weaker.

A detailed methodological objection may have been produced by an experienced specialist. It may have been produced by an adjacent-domain researcher working carefully with AI assistance. It may have been produced by a novice using AI responsibly and verifying each step. Or it may be a fluent but fundamentally mistaken objection generated through superficial prompting. The surface form of the criticism increasingly reveals less about the competence behind it. This makes both pure credentialism and pure content egalitarianism inadequate responses.

Pure credentialism would assign credibility primarily according to the status of the critic. This remains efficient and often rational because credentials and disciplinary experience contain information about expertise. But it creates false negatives: technically valid criticism may be discounted because its source lacks conventional authority.

Pure content egalitarianism would treat all criticisms as presumptively equivalent until their substance has been independently evaluated. This avoids status exclusion but creates a severe scaling problem. If AI lowers the cost of generating plausible technical objections, scientific institutions cannot realistically subject every objection to full expert adjudication. The governance problem is therefore one of *credibility allocation under conditions of cheap criticism*. One possible response is to place greater weight on properties of the criticism that can be independently inspected. These properties provide evidence about the credibility of the particular criticism. This approach does not ask only: *Who is making the claim?*

It also asks:

- Is the criticism tied to identifiable evidence?

- Can the cited source be retrieved and checked?

- Are analytical steps disclosed?

- Can another evaluator reproduce the reported discrepancy?

- Are assumptions stated explicitly?

- Does the criticism survive correction of obvious misunderstandings?

- Is uncertainty acknowledged?

- Can the critic specify what evidence would refute the objection?

- Is the criticism responsive to counterevidence?

These criteria shift part of the evaluation from personal status toward demonstrable epistemic performance.

They do not replace expertise. Indeed, many questions cannot be settled without specialist tacit knowledge. A reproducible calculation may still be irrelevant to the substantive question. A statistically correct alternative specification may be scientifically inappropriate. Domain experience may reveal constraints invisible to a general-purpose AI system or an external critic. Such indicators therefore operate alongside, rather than instead of, traditional expertise indicators. The resulting model is additive rather than substitutive:

status-based credibility + independently inspectable evidence → assessment of a scientific challenge

Status provides prior information about likely competence. Performance provides evidence about the quality of the particular criticism. The relative weight of each may differ across cases.

An anonymous claim stating only that a result “looks wrong” provides little basis for scientific action. An anonymous commenter who identifies duplicated images, provides exact figure locations, documents the comparison procedure, and supplies independently inspectable evidence presents a different case. Conversely, a highly credentialed scientist who offers a criticism outside the pertinent domain without evidence should not receive unlimited epistemic weight merely because of institutional status. This distinction is particularly relevant to AI-assisted scrutiny because generative AI simultaneously strengthens and weakens performance signals. It strengthens them by making it easier to produce transparent supporting materials: executable code, citation tables, alternative model specifications, audit trails, structured comparisons, and reproducible workflows.

But it weakens them because AI can also manufacture the appearance of rigor. Fluent methodological language, extensive references, code, and statistical terminology can be generated without genuine understanding. What once functioned as a costly signal of technical competence may therefore become easier to counterfeit. The important signal consequently shifts again—from technical sophistication itself toward *verifiability*. A criticism should receive greater weight not simply because it looks technically sophisticated, but because its evidentiary chain can be independently inspected.

This distinction offers a possible answer to the question raised by AI-assisted distributed scrutiny. Scientific institutions need not decide in advance whether a critic possesses sufficient social status to speak. Nor must they treat every criticism equally. They can instead distinguish between the right to raise a challenge and the evidentiary burden required for that challenge to command institutional attention. This leads to the fourth proposition:

P4 — Capability–Recognition Mismatch. The diffusion of scientific evaluative capability need not be accompanied by an equivalent diffusion of institutional recognition. As AI lowers the cost of substantive scrutiny outside formal evaluative roles, scientific institutions will increasingly need mechanisms for evaluating criticism that combine indicators of expertise with independently inspectable indicators of evidentiary performance.

The entry of anonymous concerns into formal processes demonstrates that recognition is possible; it does not establish that the mismatch has disappeared. Here, the mismatch concerns the passage from an externally raised concern to an accountable decision about its evidentiary weight. Receiving a comment, opening an investigation, establishing an error, and requiring a response are distinct institutional acts. Caron et al. (2025) document processing burdens, not the general adequacy of the thresholds connecting those acts. The remaining questions are which specific objections warrant investigation, who must assess and respond to them, and how the outcome is attributed to evidence rather than the critic's status. Existing routes can resolve individual cases while leaving those questions unsettled at scale. P4 therefore concerns the conditions under which external capability receives appropriate institutional uptake, not the absence of all recognition mechanisms.

A scientific challenge does not become valid or invalid because it comes from outside an institution. Its institutional weight must depend on whether the objection survives scrutiny of its own. The capability–recognition mismatch therefore leads to a quality-control question: how can institutions distinguish credible external scrutiny from criticism that merely appears technically sophisticated? Section 6 examines why that distinction becomes harder as criticism becomes cheap, and what it requires.

6. Risks and Governance: When Criticism Becomes Cheap

Generative AI can help critics identify errors and help institutions check objections. The directional neutrality described in Section 4.4 applies to these technical operations on both sides. It does not imply equal reductions in the total cost of generating a candidate criticism and reaching a warranted institutional decision about it.

The difference follows from the tasks each endpoint requires. Producing a candidate objection can end with a plausible claim, supporting code, or a proposed alternative analysis. Deciding whether that objection warrants correction or another institutional consequence additionally requires checking its relevance, resolving domain-specific ambiguity, considering an author's response, and assigning responsibility for the decision. AI can assist these activities, but producing an output does not itself discharge their epistemic and procedural requirements. Tacit knowledge remains necessary where contextual judgment cannot be reduced to a mechanical check; an opportunity to respond and an accountable decision remain necessary where institutional consequences are at stake. Existing publication-ethics procedures already distinguish raising a concern from determining its consequences (Committee on Publication Ethics [COPE], 2008; Wager & Kleinert, 2021).

Consequently, where these residual adjudicative tasks remain substantial, *the cost of generating a criticism can fall faster than the total cost of verifying and adjudicating it*. This is a conditional asymmetry, not a claim that AI benefits generation alone. For a readily checkable numerical discrepancy, verification can become equally cheap or cheaper. The scaling problem arises when a growing supply of candidate objections requires contextual assessment and accountable decisions that do not accelerate at the same rate.

6.1. Pseudo-Scrutiny

I use *pseudo-scrutiny* to refer to criticism that has the external form of scientific evaluation but lacks a sufficiently reliable evidentiary or analytical basis. Pseudo-scrutiny need not be deliberately deceptive. It can arise from genuine misunderstanding. A user may ask a language model to identify weaknesses in a paper and receive a technically plausible objection to an estimator that was in fact appropriately used. A model may recommend a sensitivity analysis whose assumptions are incompatible with the study design. It may identify a supposed contradiction between two papers that actually use different estimands or populations. It may describe a legitimate methodological debate as though one side has been definitively resolved.

Or it may invent a citation supporting an objection that no published source actually supports. The rhetorical quality of generative AI makes these failures particularly consequential. Weak criticism does not necessarily look weak.

Research on generative AI in evidence synthesis provides a useful warning. A systematic review of 19 studies found substantial errors across several research tasks. Generative AI systems missed 68–96% of relevant studies in literature searching; median error rates were 28% for incorrect exclusion decisions, 14% for data extraction, and 27% for risk-of-bias assessment. The authors concluded that current evidence does not support the use of generative AI for evidence synthesis without human involvement or oversight (Clark et al., 2025). These findings concern evidence synthesis rather than scientific criticism directly, but the implication transfers. Many of the operations required to criticize a scientific paper—literature retrieval, evidence comparison, data extraction, and methodological classification—are themselves vulnerable to AI error.

The existence of an AI-assisted objection therefore provides evidence that scrutiny has occurred. It does not provide evidence that the scrutiny was competent.

6.2. Automation Bias and Epistemic Dependence

A second risk is not that AI produces an error, but that the human evaluator fails to detect it. Automation bias refers broadly to the tendency to over-rely on automated recommendations or outputs. Research on human–AI collaboration identifies over-trust, attentional constraints, AI literacy, professional expertise, task-verification demands, and explanation complexity as factors influencing such reliance (Romeo & Conti, 2026). Experimental work with generative AI likewise shows that users supplied with incorrect AI assistance can perform worse than users receiving no assistance. Warning interventions can mitigate this effect without eliminating it (Wingerter et al., 2025). This matters because AI-assisted distributed scrutiny depends on a human–AI relationship that is structurally different from conventional expert use of software.

Traditional statistical software usually executes a procedure explicitly specified by the user. A generative model can instead propose the procedure, explain why it is appropriate, generate the code, interpret the result, and formulate the criticism.

The same system may therefore provide both the analysis and the justification for trusting the analysis. That creates the possibility of *epistemic dependence*: a critic appears to be independently evaluating a scientific claim while relying on the same AI system for the choice of objection, the technical reasoning behind it, and the interpretation of the evidence. The danger is not merely that the model may be wrong. It is that the user may lack the independent competence required to recognize where it is wrong. The resulting expansion of

evaluative capability can therefore be shallow. A person may acquire the ability to *perform* an analytical operation without acquiring the ability to independently judge when that operation is appropriate.

This distinction is especially important for the argument developed in this paper. Lowering a capability threshold is not identical to eliminating the expertise gradient above that threshold.

6.3. Hallucinated Criticism

Citation hallucination illustrates the problem in its clearest form. Generative systems can produce bibliographic references that appear legitimate but do not correspond to real publications, corrupt author names or titles, or attach genuine references to claims those references do not support. Hallucinated references are no longer confined to laboratory demonstrations of LLM behavior. Sakai et al. (2026) identified approximately 300 papers containing at least one hallucinated citation across proceedings of the Association for Computational Linguistics, its North American Chapter, and Empirical Methods in Natural Language Processing conferences from 2024–2025. The same mechanism can operate in criticism. A critic might challenge a published claim using an apparently authoritative counter-reference that does not exist.

A model might attribute a methodological position to a real paper that never made it. A literature comparison may therefore look extensively sourced while resting on fabricated or semantically mismatched evidence. This creates a peculiar recursive problem. AI may be used to detect hallucinated references in a scientific paper. AI may also generate hallucinated references in the criticism of that paper. Scientific evaluation thus acquires another layer:

claim → criticism of claim → verification of criticism

The cost reduction generated by AI can occur at each layer, but so can error generation.

6.4. Criticism Flooding

The problem becomes more serious when multiplied by scale. Conventional criticism has historically been expensive. Writing a detailed methodological response, reproducing an analysis, or tracing dozens of references requires substantial time. These costs limit participation but also constrain volume. Generative AI weakens that volume constraint. A single motivated user may be able to generate dozens of detailed objections to a paper. An organized group may be able to scrutinize hundreds of publications. Automated agents could potentially produce review-like evaluations continuously. The resulting increase is not necessarily malicious. Much of it may reflect legitimate attempts to improve science. But scientific institutions face a finite verification capacity.

Editors have limited time. Authors have limited capacity to respond. Research-integrity offices have limited investigative resources. Statistical reviewers and subject-matter experts remain scarce. If the marginal cost of producing a criticism approaches zero while the marginal cost of adjudicating it remains high, institutions may face a *criticism-flooding problem*.

Post-publication peer review already provides an early analogue. PubPeer has enabled the identification of consequential problems in published research, but the number, frequency, and velocity of comments raising possible integrity concerns have also created substantial administrative burdens for institutions required to determine which allegations warrant investigation (Caron et al., 2025). Generative AI could amplify this existing asymmetry by lowering the cost of producing technically elaborate challenges without proportionately reducing the cost of adjudicating them. The analogy is not to spam because scientific

criticism should not be presumed worthless. The relevant point is structural: when submission costs fall dramatically while adjudication costs do not, screening becomes a central governance problem.

The problem is therefore not excessive criticism as such. It is a mismatch between *criticism-generation capacity and criticism-verification capacity*.

6.5. Provenance and Auditability

One response is to make AI-assisted criticism more auditable. A criticism that relies on generative AI should ideally expose enough of its evidentiary pathway for others to inspect the result without trusting either the critic or the AI system. This suggests several desirable properties:

- identifiable source documents;

- retrievable quotations or claim locations;

- disclosed analytical code;

- reproducible transformations;

- explicit model assumptions;

- versioned data and software where applicable;

- clear separation between retrieved evidence and AI-generated interpretation;

- disclosure of substantial AI assistance where relevant;

- correction mechanisms when AI-generated errors are identified.

These properties can be understood through *provenance*. The Findable, Accessible, Interoperable, and Reusable (FAIR) principles include provenance among the requirements supporting reuse, and the World Wide Web Consortium (W3C) PROV framework describes the entities, activities, and agents involved in producing digital objects (Wilkinson et al., 2016; Groth & Moreau, 2013). The central question is not merely whether AI was used. It is whether another evaluator can determine where the criticism came from, which evidence supports it, which transformations were performed, and which parts depend on generative inference. This is particularly important because demanding complete disclosure of every prompt or interaction may be impractical and epistemically unhelpful. A long conversational transcript does not necessarily make a criticism reproducible.

What matters more is the provenance of the substantive evidentiary chain. If the objection concerns a citation mismatch, the original claim and cited source should be identifiable. If it concerns a reanalysis, the data, code, and analytical specification should be inspectable. If it concerns a literature comparison, the inclusion process and underlying studies should be traceable. The governance objective should therefore be *auditability rather than performative transparency*.

6.6. From Technical Sophistication to Verifiability

Generative AI makes technical sophistication easier to produce, weakening its value as a signal of epistemic quality. Complex statistical language, extensive citations, and polished code deserve weight only insofar as their evidentiary basis survives independent checking. Verifiability, provenance, and reproducibility therefore become more informative than technical appearance alone. This is the credibility implication of Section 5: reducing surface-level defects can leave, or amplify, substantive errors concealed by a convincing presentation.

6.7. Governance Principles for Distributed Scrutiny

The preceding risks establish the requirements stated below as P5. Access to challenge should remain open, while institutional attention should depend on evidentiary specificity, reproducibility, and relevance. Disclosure of AI assistance cannot establish an objection's correctness. Expert adjudication remains necessary where technical checks leave contextual or methodological questions unresolved. These requirements distinguish detecting a possible anomaly from establishing an error or misconduct. Section 7 develops their institutional implementation in a single layered architecture.

6.8. The Scrutiny Quality Problem

This leads to the fifth proposition:

P5 — Scrutiny Quality Problem. As generative AI lowers the cost of producing scientific criticism beyond established evaluative roles, institutional recognition cannot rely on AI assistance or technical appearance alone. The credibility of distributed scrutiny therefore depends increasingly on provenance, reproducibility, evidentiary specificity, and appropriate expert adjudication.

The proposition captures a central paradox of scientific democratization. Lower barriers are valuable because they allow more potential errors to be noticed. Lower barriers are dangerous because they also allow more false errors to be alleged. The solution is not to restore the old barriers merely because they filtered noise. Nor is it to assume that participation itself guarantees epistemic improvement. The governance problem is to preserve the expanded capacity for challenge while raising the quality of the mechanisms through which challenges acquire scientific weight. *Criticism is becoming cheaper. Credible criticism is not.* Or, more precisely, the scarce resource increasingly becomes the capacity to verify, prioritize, and adjudicate criticism at the rate at which it can be produced.

7. Implications for Scientific Institutions

The dual-democratization model has implications beyond the use of generative AI by individual researchers. If evaluative capability becomes more widely distributed, then scientific institutions face a problem that cannot be solved solely by regulating AI use within existing expert roles. The institutional question becomes broader: *How should science respond when the capacity to identify potentially important problems spreads more widely than the authority traditionally granted to those who identify them?* The answer should not be to recreate conventional credential barriers around criticism. Nor should institutions treat every objection as epistemically equivalent. The more plausible direction is a system that separates *access to challenge* from *entitlement to institutional attention*.

Anyone may be able to raise a scientific objection. Increasing levels of institutional consequence, however, should require increasing levels of evidentiary support, reproducibility, and verification. This distinction has

implications for journals, post-publication review, research institutions, Open Science policy, and the role of non-traditional evaluators.

7.1. Journals: From Reviewer Identity to Verifiable Objection

Scientific journals have historically organized criticism largely through recognized roles. Editors select reviewers. Reviewers evaluate manuscripts. Authors respond. Readers may later submit correspondence, commentaries, or post-publication critiques, but these mechanisms usually occupy a secondary position relative to pre-publication peer review (Hardwicke et al., 2022). Generative AI may make this architecture increasingly incomplete. A reader who was not selected as a reviewer may nevertheless identify a citation error, reproduce an analysis, detect an inconsistency between a table and underlying data, or discover that a methodological assumption is unsupported. The relevant question for a journal should therefore not be only: *Who is making the criticism?*

It should increasingly also be: *Can the criticism be independently checked?* This does not imply that reviewer expertise becomes unimportant. Expert judgment remains essential for many disputes, particularly where evaluation depends on tacit knowledge, contested methodological standards, or substantive interpretation. But expertise and verifiability serve different functions. Expertise helps institutions evaluate difficult claims. Verifiability helps institutions determine whether a criticism deserves to reach that stage. Journals could therefore benefit from treating post-publication objections through structured evidentiary categories rather than through a binary distinction between formal reviewer and outside critic. For example, a claim of numerical inconsistency might require the relevant table, calculation, and reproducible derivation.

A citation challenge might require the disputed statement, the cited source, and the precise mismatch. A reanalysis might require code, data provenance, analytical specification, and enough information to reproduce the result. A conceptual or methodological objection might require explicit identification of the disputed assumption and relevant literature. Such requirements would not determine whether the criticism is correct. They would determine whether it is sufficiently specified to warrant further attention. The resulting principle is: *low barriers to raising an objection, higher evidentiary thresholds for escalating its institutional consequences.*

7.2. Post-Publication Peer Review: Designing for Scale

Post-publication peer review is particularly important under distributed scrutiny because it provides a route through which evaluations originating outside conventional review roles can become visible (PubPeer, n.d.). Its importance may therefore increase. So may its governance burden. If AI lowers the cost of identifying, formulating, and submitting potential problems, post-publication systems may receive more criticism than existing editorial and research-integrity structures can feasibly adjudicate. The appropriate response is not to discourage post-publication criticism. Doing so would sacrifice one of the primary benefits of expanded scrutiny. Instead, systems may need to distinguish among levels of claim. A useful architecture would separate:

observation from *evidence-linked anomaly* from *demonstrated error* from *allegation of misconduct*. These categories are not interchangeable. An unexplained discrepancy may warrant attention without demonstrating that a paper is invalid. A reproducible analytical error may undermine a result without implying misconduct. A suspicious pattern may justify investigation without itself establishing intent. Generative AI makes this distinction more important because it can produce rhetorically forceful allegations before the underlying evidence has been adequately established. Post-publication review systems should

therefore reward specificity over accusation. A short criticism that identifies a reproducible error may be more valuable than a lengthy AI-generated review containing dozens of speculative concerns.

The unit of attention should increasingly become the *verifiable objection*, not the volume or technical sophistication of the critique.

7.3. Research Institutions: Triage Before Investigation

Universities, hospitals, laboratories, funders, and research-integrity offices face a related challenge. These institutions cannot investigate every externally submitted criticism as though it constituted a formal allegation of misconduct. Yet dismissing objections solely because they originate outside recognized professional networks may cause genuine problems to be missed. Distributed scrutiny therefore creates a need for better triage. A possible institutional sequence is:

receipt of challenge → evidentiary sufficiency check → technical reproduction where feasible → domain review → research-integrity investigation only where warranted

This sequencing matters. It prevents the act of raising a technical anomaly from being conflated with an accusation of wrongdoing. It also prevents institutional status from becoming the sole gateway through which anomalies can receive attention. Under such a model, an outsider need not establish misconduct. The initial burden is narrower: provide enough evidence that a specific scientific concern can be independently examined. The institution then bears responsibility for determining whether the concern survives technical verification and whether further action is justified. This division of labor may become increasingly important as criticism-generation capacity grows faster than formal investigative capacity.

7.4. Open Science

The implications also extend to Open Science. Open Science and FAIR articulate goals for accessible and reusable research materials (UNESCO, 2021; Wilkinson et al., 2016). Making publications, data, and code accessible remains essential, but access alone does not ensure that scientific claims can be meaningfully examined by a wider range of actors. The access–capability distinction developed earlier in this article therefore matters at the level of scientific infrastructure as well as individual competence.

Generative AI may reduce some of the technical and cognitive barriers that remain after access has been granted. Yet its contribution to scrutiny depends on whether scientific objects are available in forms that permit claims to be traced, calculations to be reconstructed, analyses to be reproduced, and disagreements to be linked to identifiable evidence. Openness should therefore support not only access and reuse, but the practical ability to trace, interrogate, reproduce, and challenge scientific claims.

This does not replace existing goals of transparency, reproducibility, or reuse. Rather, distributed scrutiny gives those goals an additional institutional significance: they determine how readily a criticism originating outside established evaluative roles can be independently examined rather than accepted or dismissed on the basis of the critic’s status.

7.5. Non-Traditional Evaluators and Scientific Reputation

Distributed scrutiny matters for independent researchers, practitioners, patient researchers, technical specialists, and others who possess relevant experience without occupying conventional evaluative roles. AI assistance can help them convert substantive concerns into inspectable objections. It does not confer equivalent domain expertise or establish that every objection warrants institutional acceptance.

The recognition problem concerns the sufficiency of credentials, not their abolition. Institutional affiliation, disciplinary standing, and prior expertise remain informative, especially when adjudication depends on tacit knowledge. They need not determine in advance whether a specific objection deserves examination. Equally, technically sophisticated presentation cannot substitute for verification simply because AI makes it easier to produce. Recognition can attach conditionally to the criticism that survives scrutiny, without granting the critic general expert standing. A system that accepts evidence-linked objections and escalates them according to their demonstrated merits preserves the benefits of broader participation while retaining the role of expertise.

7.6. A Layered Governance Architecture

Taken together, these implications support a layered governance model for AI-assisted distributed scrutiny. Its procedural elements have precedents. COPE's flowchart for suspected fabricated data in a published article moves from a reader's concern to assessment, contact with the author, and further action according to the response (COPE, 2008). The subsequent CLUE recommendations distinguish journals' responsibility for the published record from institutions' misconduct investigations and explicitly cover concerns beyond misconduct (Wager & Kleinert, 2021). The following layers do not claim to invent research-integrity triage. Their purpose is to connect such procedures to a wider intake of AI-assisted objections, including ordinary analytical and interpretive challenges that do not allege misconduct, which existing integrity guidance already recognizes. The contribution is the relationship between expanded evaluative capability and the evidentiary thresholds for institutional uptake, rather than a new misconduct-investigation protocol.

Layer 1: Open Challenge

Scientific objections should be relatively easy to raise regardless of the critic's institutional status. At this stage, the purpose is anomaly detection, not adjudication.

Layer 2: Evidence Linking

A criticism seeking institutional attention should identify the relevant claim, evidence, data, analysis, or source with sufficient precision to permit checking.

Layer 3: Reproducibility and Technical Verification

Where the criticism involves a calculation, citation, dataset, or analytical result, the objection should be independently reproduced where feasible.

Layer 4: Domain Adjudication

Questions that cannot be resolved mechanically should be evaluated by people with appropriate substantive or methodological expertise.

Layer 5: Institutional Consequence

Correction, expression of concern, retraction, rebuttal, dismissal, or research-integrity investigation should occur only after the preceding levels justify escalation.

Citizen science offers an adjacent precedent for managing distributed input through layered verification. Baker et al. (2021) reviewed 259 ecological citizen-science schemes and found verification information for 142. They describe expert, community, and automated approaches and propose hierarchical verification that

escalates flagged records to experts. Scientific criticism differs from species-record verification, but the institutional principle is informative: broad participation need not imply uniform verification at every stage. This architecture separates participation from adjudication. It allows broad participation at the point where broad participation is most valuable: discovering possible problems. It becomes increasingly selective as the consequences of the criticism increase.

This principle can be expressed as: *openness at the point of detection, rigor at the point of consequence.*

7.7. From Peer Review to a Wider Verification Ecology

Conventional peer review remains a central institution for expert judgment within a wider *verification ecology*. Readers and independent analysts examine published results; computational systems assist anomaly detection; domain experts assess ambiguous findings; editors and institutions determine warranted consequences. Generative AI changes the distribution of capability among these actors. The question is whether their combined work converts expanded scrutiny into reliable scientific correction. Where potential criticism grows faster than accountable adjudication, broader access to challenge must be accompanied by mechanisms for prioritizing evidence and resolving disputes. The layered architecture locates those requirements without assigning every evaluative task to formal peer reviewers.

7.8. Scope Conditions and Limitations

This is a conceptual account of a mechanism and its institutional implications, not an estimate of its prevalence or net effect. Its scope requires usable access to scientific materials, AI assistance appropriate to the task, and users able to check at least the evidentiary steps they report. Lower task costs do not establish that access is equitable, that evaluation improves across all disciplines, or that institutions receive a larger proportion of valid objections.

P3 has an empirical implication beyond increased participation: under specified tasks and access conditions, AI assistance should enable actors outside assigned evaluative roles to produce more independently validated components of scrutiny, or to produce them at lower cost. Comparisons should hold task difficulty and available materials constant and assess objections independently of their source. More fluent criticism without improved evidentiary performance would not support P3. Repeated failures to improve validated performance or reduce its cost across the stated conditions would count against the proposed mechanism in those settings.

P4 would gain support if independently validated objections from outside recognized evaluative roles receive less timely or less substantive institutional consideration than comparably relevant and well-supported objections from established evaluators. Such comparisons must distinguish source status from differences in evidence, relevance, and procedural requirements. Consistently equivalent, timely, and evidence-sensitive uptake across these groups would count against the proposed mismatch in the institutions studied.

P5 combines an empirical rationale with a normative recommendation. Its empirical rationale gains support if provenance, reproducibility, and evidentiary specificity help distinguish valid from invalid objections beyond AI-use disclosure or technical appearance, and if appropriate expert adjudication improves decisions where technical checks are insufficient. If these features add no discriminatory value, or their burdens systematically exclude valid challenges without improving decisions, the proposed implementation requires

revision. The normative recommendation also depends on the value assigned to open challenge and accountable institutional action; observational results alone cannot settle that commitment.

Finally, the cost asymmetry is conditional. If verification and accountable adjudication scale at least as fast as candidate criticism, the predicted bottleneck will not arise. That result would challenge the scaling argument without refuting the broader possibility of distributed scrutiny.

8. Conclusion

Generative AI can lower barriers to interrogating scientific claims as well as producing them. The access–capability distinction explains why this matters: access to an article, dataset, or analytical method does not by itself confer the ability to evaluate it. AI assistance can reduce some of the technical costs that remain after access has been achieved. This is the scrutiny component of *dual democratization*, specified in P3 as the expansion of substantive evaluative tasks beyond assigned roles.

That expansion creates a *capability–recognition mismatch* where institutions' routes for receiving and weighting external criticism fail to keep pace with evaluative capability. Existing post-publication mechanisms demonstrate that such recognition is possible, while leaving questions about its thresholds, responsibilities, and scale. Lowering capability barriers also lowers barriers to plausible but invalid objections. The resulting *scrutiny quality problem* concerns how to identify criticism that deserves institutional action.

The governing principle is *openness at the point of detection, rigor at the point of consequence*. Criticism should identify a specific claim, connect it to inspectable evidence, survive technical verification where feasible, and receive appropriate domain adjudication where necessary. Recognition can then attach conditionally to the criticism that survives scrutiny. This retains the role of expertise while allowing non-traditional evaluators to contribute to scientific correction.

As access and some capability barriers decline, the practical bottleneck may increasingly shift toward verification, prioritization, and adjudication. Scientific institutions may therefore face a different allocation problem: not simply how to enable criticism, but how to identify which objections warrant further attention and what level of evidence should be required before increasingly consequential action is taken. The central question for science in an AI-mediated environment may therefore become less: *Who is qualified to criticize?* and increasingly: *What must a criticism demonstrate before science should act on it?* Generative AI may democratize science not only by changing who can produce scientific claims.

It may also change who can challenge them. Whether that transition improves scientific self-correction will depend on whether institutions can preserve openness to challenge while making credibility depend on evidence that survives verification.

References

Alvarez, C., Bennett, M., & Wang, L. L. (2024). Zero-shot Scientific Claim Verification Using LLMs and Citation Text. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)* (pp. 269–276). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.sdp-1.25

Baker, E., Drury, J. P., Judge, J., Roy, D. B., Smith, G. C., & Stephens, P. A. (2021). The Verification of Ecological Citizen Science Data: Current Approaches and Future Possibilities. *Citizen Science: Theory and Practice, 6*(1), Article 12. https://doi.org/10.5334/cstp.351

Caron, M. M., Lye, C. T., Bierer, B. E., & Barnes, M. (2025). The PubPeer conundrum: Administrative challenges in research misconduct proceedings. *Accountability in Research, 32*(8), 1369–1387. https://doi.org/10.1080/08989621.2024.2390007

Clark, J., Barton, B., Albarqouni, L., Byambasuren, O., Jowsey, T., Keogh, J., Liang, T., Moro, C., O'Neill, H., & Jones, M. (2025). Generative artificial intelligence use in evidence synthesis: A systematic review. *Research Synthesis Methods, 16*(4), 601–619. https://doi.org/10.1017/rsm.2025.16

Committee on Publication Ethics. (2008). *What to do if you suspect fabricated data: (b) Suspected fabricated data in a published article* [Flowchart]. https://content.ifzg.hr/poveznice/Ethical_codex_for_Editors-COPE-english.pdf

Croce, M., & Marsili, N. (2025). Misplaced Trust in Expertise: Pseudo-Experts and Unreliable Experts. *Social Epistemology, 39*(6), 667–682. https://doi.org/10.1080/02691728.2025.2491104

Daepp, M. I. G., Tomlinson, K., Counts, S., & Suri, S. (2026). AI and the democratization of knowledge work. *Nature Computational Science, 6*(6), 558–564. https://doi.org/10.1038/s43588-026-00985-z

Edelman, B., & Skolnick, J. (2025). Valsci: an open-source, self-hostable literature review utility for automated large-batch scientific claim verification using large language models. *BMC Bioinformatics, 26*(1), Article 140. https://doi.org/10.1186/s12859-025-06159-4

Groth, P., & Moreau, L. (Eds.). (2013). *PROV-Overview*. W3C Working Group Note. https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

Hao, Q., Xu, F., Li, Y., & Evans, J. (2026). Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature, 649*(8099), 1237–1243. https://doi.org/10.1038/s41586-025-09922-y

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Tzavella, L., Bendixen, T., Handcock, S. A., Köneke, V. E., & Ioannidis, J. P. A. (2022). Post-publication critique at top-ranked journals across scientific disciplines: a cross-sectional assessment of policies and practice. *Royal Society Open Science, 9*(8), Article 220139. https://doi.org/10.1098/rsos.220139

Keren, A. (2025). Expert Authority and Its Assessment. *Social Epistemology, 39*(6), 612–625. https://doi.org/10.1080/02691728.2025.2472780

Markowitz, D. M. (2024). From complexity to clarity: How AI enhances perceptions of scientists and the public's understanding of science. *PNAS Nexus, 3*(9), Article pgae387. https://doi.org/10.1093/pnasnexus/pgae387

Mirkin, J., & Sojka, M. M. (2026). From Expertise to Trustworthiness: What Laypeople Can Really Assess about Experts. *Episteme*, 1–18. https://doi.org/10.1017/epi.2026.10118

Nabavi, A., Safari, F., Shmoury, A. H., Tabet, S., Perdomo-Luna, C., & Celi, L. A. (2026). Artificial intelligence in scholarly peer review: a scoping review of applications, risks, and governance challenges. *International Journal of Medical Informatics, 214*, Article 106418. https://doi.org/10.1016/j.ijmedinf.2026.106418

PubPeer. (n.d.). *Frequently asked questions*. Retrieved September 5, 2026, from https://pubpeer.com/static/faq

Rebitschek, F. G., Carella, A., Kohlrausch-Pazin, S., Zitzmann, M., Steckelberg, A., & Wilhelm, C. (2025). Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information. *npj Digital Medicine, 8*(1), Article 343. https://doi.org/10.1038/s41746-025-01752-6

Reddy, C. K., & Shojaee, P. (2025). Towards Scientific Discovery with Generative AI: Progress, Opportunities, and Challenges. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(27), 28601–28609. https://doi.org/10.1609/aaai.v39i27.35084

Rolin, K. (2025). Expert trustworthiness and the value-free ideal of science. *European Journal for Philosophy of Science, 15*(4), Article 77. https://doi.org/10.1007/s13194-025-00704-x

Romeo, G., & Conti, D. (2026). Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *AI & Society*, *41*(1), 259–278. https://doi.org/10.1007/s00146-025-02422-7

Ross-Hellauer, T. (2017). What is open peer review? A systematic review. *F1000Research, 6*, Article 588. https://doi.org/10.12688/f1000research.11369.2

Sakai, Y., Kamigaito, H., & Watanabe, T. (2026). HalluCitation Matters: Revealing the Impact of Hallucinated References with 300 Hallucinated Papers in ACL Conferences. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 47295–47376). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-long.2189

UNESCO. (2021). *UNESCO recommendation on open science*. https://www.unesco.org/en/legal-affairs/recommendation-open-science

Wager, E., & Kleinert, S. (2021). Cooperation & Liaison between Universities & Editors (CLUE): Recommendations on best practice. *Research Integrity and Peer Review, 6*, Article 6. https://doi.org/10.1186/s41073-021-00109-3

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), Article 160018. https://doi.org/10.1038/sdata.2016.18

Wingerter, T. L., Straub, T., & Schweitzer, S. (2025). Mitigating Automation Bias in Generative AI Through Nudges: A Cognitive Reflection Test Study. *Procedia Computer Science, 270*, 2106–2114. https://doi.org/10.1016/j.procs.2025.09.331

Zhang, Y., Khan, S. A., Mahmud, A., Yang, H., Lavin, A., Levin, M., Frey, J., Dunnmon, J., Evans, J., Bundy, A., Dzeroski, S., Tegner, J., & Zenil, H. (2025). Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence, 1*(1), Article 14. https://doi.org/10.1038/s44387-025-00019-5

Statements and Declarations

Funding

No funds, grants, or other support were received during the preparation of this manuscript.

Competing interests

The author has no relevant financial or non-financial interests to disclose.

Data availability

No datasets were generated or analysed for this conceptual article.

Use of generative AI

Generative AI tools were used during drafting, revision, literature identification, and manuscript integration. AI-generated suggestions and bibliographic information were checked during preparation. Responsibility for the arguments, source interpretation, and final text rests with the author.