

# Who Pays the Cost of Verification? The Economy of Judgment in AI-Mediated Scientific Production

*A theory-application conceptual article with a single documented case*

**Working byline:** Andrea Viliotti and Framework GDE  
**Andrea Viliotti:** Independent researcher, Trento, Italy  
**Framework GDE:** Workflow/public project label; see author note for venue-specific normalization  
**Keywords:** economy of judgment; AI-assisted science; verification costs; interpretive authority; large language models; epistemic infrastructure; cognitive habitat; scientific production; sophisticated wrongness; provenance  
**Submission note:** Venue-neutral manuscript prepared for interdisciplinary STS / AI & Society / digital society outlets. Correspondence details can be inserted at submission stage.

## Abstract

AI-mediated scientific production is increasingly discussed either as a productivity breakthrough or as an epistemic risk. This paper argues that neither frame is sufficient on its own because both under-specify the institutional ecology through which scientific claims are produced, verified, contested, and archived. Building on prior work that reconstructs cognitive rights as habitat rights and introduces the economy of judgment as a three-layer mechanism operating through six diagnostic axes, the article applies that framework to a single documented case: the public GDE emergent-gravity branch and its two companion manuscripts. The claim is intentionally narrow. The paper does not adjudicate the underlying physics and does not offer causal measurement or validated metrics. It asks whether the framework helps explain a structural redistribution in which production costs decrease while verification costs are shifted downstream onto disciplinary communities, archival infrastructures, and other institutions of scientific judgment. The analysis shows that the six axes—time of interpretation, visibility, provenance, capability, contestability, and institutional pluralism—map onto the scientific production chain without ad hoc modification. It also introduces the concept of **distributed capability deficit**, a condition in which no participant in the production chain can independently certify the output, and links it to the risk of **sophisticated wrongness**. Four middle-range propositions, three framework-level hard nulls, and a portability agenda are formulated to make the argument falsifiable. The contribution is diagnostic rather than prescriptive: it offers an analytically portable way to study AI-assisted scientific production across domains while keeping self-reference, provenance opacity, and external-validation limits explicit.

## Abstract in Italian

La produzione scientifica mediata dall'IA viene oggi letta soprattutto in due modi: come accelerazione produttiva o come rischio epistemico. Questo articolo sostiene che nessuna delle due cornici basti da sola, perché entrambe sotto-specificano l'ecologia istituzionale entro cui le claim scientifiche vengono prodotte, verificate, contestate e archiviate. Basandosi su due lavori precedenti — uno che ricostruisce i diritti cognitivi come diritti di habitat e uno che introduce l'economia del giudizio come meccanismo a tre layer operante attraverso sei assi diagnostici — il saggio applica quel framework a un singolo caso documentato: il ramo pubblico GDE di gravità emergente e i suoi due companion manuscripts. La tesi è intenzionalmente stretta. Il paper non decide sulla correttezza della fisica sottostante e non offre misurazione causale né metriche validate. Chiede piuttosto se il framework aiuti a spiegare una redistribuzione strutturale in cui i costi di produzione diminuiscono mentre i costi di verifica vengono spostati a valle sulle comunità disciplinari, sulle infrastrutture archivistiche e sulle altre istituzioni del giudizio scientifico. L'analisi mostra che i sei assi — tempo di interpretazione, visibilità, provenienza, capability, contestabilità e pluralismo istituzionale — si applicano alla catena di produzione scientifica senza modifiche ad hoc. Il saggio introduce inoltre il concetto di **distributed capability deficit**, cioè una condizione in cui nessun partecipante della catena di produzione può certificare autonomamente

l'output, e lo collega al rischio di **sophisticated wrongness**. Quattro proposizioni di medio raggio, tre hard nulls a livello di framework e una agenda di portabilità rendono l'argomento falsificabile. Il contributo è diagnostico, non prescrittivo: offre un modo analiticamente portabile per studiare la produzione scientifica assistita da IA in domini diversi, mantenendo espliciti i limiti di auto-riferimento, opacità di provenienza e validazione esterna.

## 1. Introduction

The question this paper addresses is not whether artificial intelligence can participate in scientific research. That question is already answered by practice: AI tools are used daily in data analysis, literature synthesis, hypothesis generation, formal derivation, manuscript drafting, and code production across virtually every scientific discipline. The harder and more consequential question is how AI mediation redistributes the conditions under which scientific knowledge is judged — verified, contested, archived, and remembered — and who bears the costs of that redistribution.

Current discussions of AI in science are dominated by two families of work that communicate poorly with each other. The first is techno-optimistic and demonstration-oriented. Romera-Paredes et al. (2024) show that LLMs can support mathematical discovery through program search. Boiko et al. (2023) push toward autonomous experimental planning in chemistry. Lu et al. (2024) propose open-ended paper generation within machine learning research. Bubeck et al. (2025) report early acceleration experiments with GPT-5. Together, these works establish that LLMs can do more than edit prose: they can participate in parts of the research loop.

The second family is cautionary and risk-oriented. Birhane et al. (2023) argue that large language models can compromise scientific trust if convenience outruns discipline. Messeri and Crockett (2024) warn against “illusions of understanding” — a regime in which output becomes more abundant and seemingly coherent while actual explanatory grip weakens. Stokel-Walker and Van Noorden (2023) survey the implications for science more broadly.

Both families are necessary. Neither provides an operational analytical framework for studying how AI mediation restructures the institutional ecology of scientific judgment. The techno-optimistic strand asks what AI can produce; it does not ask who verifies the product. The cautionary strand names the risks; it does not decompose the mechanism through which verification costs are redistributed across producers, communities, and archival infrastructures.

This paper proposes that such a framework already exists, developed for a different but structurally isomorphic domain. The economy of judgment (Viliotti 2026b) describes how AI-mediated environments redistribute interpretive authority and the social costs of verification through three interlocking layers — allocation, epistemic traceability, and institutional correction — operating along six diagnostic axes: time of interpretation, visibility, provenance, capability, contestability, and institutional pluralism. The framework was originally formulated for the governance of platforms, educational institutions, memory institutions, and public infrastructures. This paper demonstrates that it applies with diagnostic precision to the production of scientific knowledge under AI mediation.

The evidence base is a single documented case, chosen not for its typicality but for its analytical extremity. Between 29 and 31 March 2026, a non-physicist independent researcher working with frontier commercial LLMs deposited on Zenodo a seven-paper programme on emergent gravity, accompanied by a comparative novelty assessment and a meta-analytic case study of the collaboration itself. The corpus is entirely public, carries explicit maturity labels, includes pre-declared hard nulls, and contains one executed negative result. It is, to the author's knowledge, the most fully documented public case of AI-assisted theoretical-physics programme construction currently available. It is also the

author's own work, which creates a self-referential structure that must be managed rather than concealed.

The paper makes three contributions. First, it shows that the economy of judgment, originally formulated for AI-mediated societies, applies to AI-mediated science without ad hoc modification: the same six axes and three layers diagnose the redistribution of verification costs in scientific production. Second, it introduces the concept of distributed capability deficit — a regime in which no participant in the production chain (neither the human architect nor the LLM mediator) can independently certify the scientific validity of the output — and connects it to the previously defined notion of sophisticated wrongness (Viliotti 2026c). Third, it formulates four middle-range propositions specific to AI-mediated science, designed to be testable against independent cases beyond the one studied here.

The paper is a theory-application conceptual article with a single documented case. It does not claim empirical generalizability from one case. It claims diagnostic portability of a framework, which is a different and weaker claim — but one that can be strengthened or falsified by future applications to independent cases.

## 2. Research design, evidence class, and case boundary

This manuscript is a theory-application conceptual article built around a single documented case. Its central claim is deliberately narrower than a general theory of AI-assisted science: it argues that the economy of judgment offers a portable diagnostic framework for analyzing how AI mediation redistributes interpretive authority and verification costs in scientific production. The paper does not claim empirical generalizability from one case. It claims diagnostic portability, which is a weaker and more falsifiable claim.

Methodologically, the article combines three layers of material that must be kept distinct. The first is the conceptual foundation supplied by two prior works: *Diritti cognitivi come diritti di habitat nell'era dell'IA* and *L'economia del giudizio nelle società mediate dall'IA*. Those works provide the habitat framing, the three-layer mechanism, the six diagnostic axes, and the prior argument that AI-mediated environments redistribute both interpretive authority and the social costs of verification. The second is the documented case corpus: five public papers from the GDE emergent-gravity branch and two companion manuscripts that assess, respectively, the branch's comparative novelty and the human–AI collaboration that produced it. The third is a selective scholarly and policy refresh on AI-assisted science, authorship, disclosure, and scientific infrastructure, used to verify citation details, DOIs, and the current framing of editorial norms.

The article therefore operates with an evidence split rather than with a single undifferentiated literature review. The scholarly corpus concerns AI-assisted science, expertise, epistemic infrastructures, and adjacent governance debates. The normative-institutional corpus concerns disclosure, authorship, provenance, and contestability standards. The case corpus concerns a bounded public artifact sequence. Maintaining this separation is methodologically important because the companion manuscripts are not external validation. They are part of the same public project and must be cited as such: visible supports for the case, not invisible guarantees of correctness.

The methodological stance is analytic rather than forensic. This distinction is load-bearing. The paper studies a public artifact sequence, not a complete interaction ledger. It can therefore speak strongly about public outputs, explicit hard nulls, maturity labels, companion dependence, and the architecture of the resulting programme. It cannot speak strongly about keystroke-level provenance, hidden discarded branches, model-specific task allocation at session resolution, or counterfactual workflows that were never frozen. Those absences are not buried caveats. They define the evidentiary boundary of the article.

This analytic-not-forensic stance follows directly from the public evidence package. The available record names project-level tool support, public outputs, dates of deposit, explicit hard nulls, and declared role asymmetries. It does not include session logs, model-specific ledgers, or frozen alternative arms. Accordingly, the paper treats provenance as partially declared but not fully reconstructible. It also treats the H/A/HA attribution matrix from the companion case study as a disciplined analytic proposal rather than as validated forensic fact. This matters because the article's argument does not require full reconstruction of the collaboration. It requires only enough evidence to show that the public production chain exhibits a particular structure: compressed production, partially opaque provenance, distributed verification burdens, and contracted institutional mediation.

The article also preserves a strict scope boundary with respect to the underlying physics. It is not a physics-validation paper and not a hidden novelty proof by social-theory means. Paper A is used as a companion manuscript for the conditional status of the non-collapse claim; Paper B is used as a companion manuscript for the analytic decomposition of the human–AI collaboration; Papers I–V are used as the public case object. None of these documents is treated as external peer validation. The social-science claim made here is conditional on that boundary being respected.

This scope discipline extends to quantification. The paper introduces no calibrated social-science model, no new ranking, and no forward-looking numerical forecast. Numbers appear only when they are already observed in the public case corpus and are necessary to characterize the evidentiary status of the case—most notably the structural coefficient ( $= 1.38$ ) and the executed negative result in the per-galaxy residual discriminator. In the social-science argument proper, GDE functions as a heuristic and interpretive language, not as a calibrated quantitative engine.

The genre is therefore best described as an abductive, transparency-oriented conceptual article with a single documented case, companion-manuscript dependence declared in the open, and a falsifiability architecture designed to make the portability claim contestable. That combination is unusual, but it is precisely what the argument requires. The paper asks not whether the branch is correct in physics, but what happens to the ecology of scientific judgment when scientifically formatted output can be produced faster than its provenance can be reconstructed and faster than disciplinary institutions can absorb the costs of verification.

### 3. Analytical framework and adjacent frames

#### 3.1 From access to judgment

The economy of judgment was introduced to describe a redistribution that adjacent frameworks capture only in part (Viliotti 2026b). The attention economy (Simon 1971) correctly names the scarcity of salience but stops before the decisive threshold: what happens after attention is captured. Epistemic welfare (Hyzen et al. 2026) focuses on the conditions and capabilities of epistemic agency but underspecifies the institutional and material infrastructure that distributes those conditions. Epistemic injustice (Fricker 2007) reveals how credibility and intelligibility are unequally distributed but does not trace the sociotechnical chain from interface to archive. Public contestability (Cohen and Suzor 2024) identifies a crucial democratic condition but treats it as one value among others rather than as one component of a relational mechanism.

The economy of judgment does not replace these frameworks. It shows their common limit: each isolates a portion of the problem. The economy of judgment follows the joint redistribution of interpretive authority and verification costs along the full chain — from interfaces, ranking, and synthesis to schools, archives, libraries, museums, documentary standards, and sites of appeal.

### 3.2 The three-layer mechanism

The framework organizes the redistribution through three interacting layers.

The allocation layer determines what reaches judgment and with what temporal compression. Time and visibility operate here: they select what becomes salient enough to demand interpretation, and how much cognitive time remains available for that interpretation.

The epistemic traceability layer determines whether the output can be reconstructed, verified, and placed within a documentary chain. Provenance and capability operate here: they condition whether the subject can trace the origin of a claim and whether they possess the competence to evaluate it critically.

The institutional correction layer determines whether interpretive authority can be appealed, redistributed, or rebalanced. Contestability and institutional pluralism operate here: they determine whether formal or informal channels exist for challenging conclusions, holding producers accountable, and maintaining a plurality of mediating institutions.

### 3.3 The six diagnostic axes

The six axes are not a checklist of desirable values. They are dimensions of a relational mechanism, derived from an iterative cross-reading of four bridge literatures (social theory of valuation and judgment devices; platform and algorithmic mediation studies; AI literacy and education; AI in archives, libraries, museums, and cultural heritage data), a subordinate dialogue with work on knowledge and epistemic infrastructures (Edwards et al. 2013), and a normative-textual corpus (Viliotti 2026b).

- (1) Time of interpretation — the cognitive time available for verification, contextualization, and revision after the production of an output.
- (2) Visibility — the salience, ranking, and accessibility of outputs within the relevant information environment.
- (3) Provenance — the reconstructibility of the documentary chain: who produced what, with what sources, through what process, and under what responsibilities.
- (4) Capability — the competences required to read, compare, dissent, and use outputs critically, understood as a habitat-dependent rather than purely individual attribute.
- (5) Contestability — the availability of documented channels for challenge, appeal, correction, and accountability.
- (6) Institutional pluralism — the plurality of mediating institutions and sites of judgment that prevent a monoculture of interpretive authority.

### 3.4 Cognitive rights as habitat rights

The prior step in the framework (Viliotti 2026a) reconstructed cognitive rights not as a closed category of positive law but as an interpretive frame for the habitat-dependent conditions of understanding, memory, and participation. The habitat is composed of schools, archives, libraries, museums, territories, digital intermediaries, families, and associations. Judgment is not defended as a private property of individual consciousness; it requires places, practices, temporalities, professions, protocols, and institutions.

This concept translates directly into the scientific domain. The disciplinary community is the cognitive habitat of scientific judgment. Journals are institutions of provenance. Peer review is an institution of contestability. Preprint servers are infrastructures of visibility. Seminars and conferences are sites of institutional pluralism. When any of these institutions is eroded, underfunded, or bypassed,

the habitat degrades — not because individual scientists become less intelligent, but because the institutional conditions of collective verification weaken.

## 4. Case and corpus: the public GDE branch and its two companions

### 4.1 The corpus

The documented case analyzed here comprises five public papers from the GDE emergent-gravity branch (Papers I–V) deposited between 29 and 31 March 2026, plus two companion manuscripts produced on 1 April 2026 (Paper A and Paper B). All seven public documents are openly accessible with DOIs.

Paper I establishes local strong hyperbolicity for the patched effective theory (Viliotti 2026d). Paper II states the Einstein window and freezes a gravitational-wave confirmatory protocol (Viliotti 2026e). Paper III provides the micro-to-EFT bridge, RG closure, and an explicit hedgehog defect (Viliotti 2026f). Paper IV extracts the structural coefficient  $\xi = 1.38$  plus/minus 0.11 from SPARC and Planck data (Viliotti and Framework GDE 2026g). Paper V executes a per-galaxy residual discriminator and reports a formally negative result (Viliotti 2026h). Paper A is a comparative novelty assessment against five neighboring programmes (Viliotti and Framework GDE 2026i). Paper B is an analytic case study of the human-AI collaboration that produced the branch (Viliotti and Framework GDE 2026c).

### 4.2 Why this case is analytically useful

The case is useful not because it is typical but because it is extreme in ways that make the economy-of-judgment mechanism maximally visible. Four features matter.

First, the human producer is not a professional theoretical physicist. Andrea Viliotti is an independent AI strategy consultant with over forty years of experience in the technology sector but no formal training or prior refereed publication in theoretical physics. This means the capability asymmetry between human and machine is structurally maximal: the human brings programme architecture but not field expertise; the LLM brings derivational throughput but not physical judgment.

Second, the LLM mediation is declared at project level. The frozen project metadata name ChatGPT Pro 5.4 and Claude Opus 4.6 as support tools. However, session-level logs, model-specific task ledgers, and counterfactual arms are not available. Provenance is therefore declared but not fully verifiable — a condition that mirrors, at the level of scientific production, the partial transparency documented in the ChatAmsterdam mini-case of the economy-of-judgment paper.

Third, the corpus contains explicit maturity labels and pre-declared hard nulls. Each paper labels its claims as theorem-level, protocol-level, executed, or negative. Paper A defines six hard nulls (HN-1 through HN-6) for the novelty claim. Paper V defines four protocol hard nulls (N1 through N4) for the residual discriminator. Paper B defines three hard nulls (HN-B1 through HN-B3) for the collaboration claim. This is contestability offered by the producer — but its activation depends on whether anyone in the disciplinary community invests the cost of verification.

Fourth, the corpus includes one executed negative result. Paper V reports that neither internal proxies nor the MOND external-field parameter correlates significantly with per-galaxy RAR residuals at the pre-declared decision rule. The programme does not hide a failed test; it serializes one. This is epistemically significant because a corpus that contains only successes is harder to distinguish from sophisticated wrongness than a corpus that exposes its own non-confirmatory outcomes.



### 4.3 What the case does not demonstrate

The case does not demonstrate that the underlying physics is correct. It does not demonstrate that the collaboration is replicable. It does not demonstrate that AI-assisted science is generally viable. It demonstrates something narrower but analytically sufficient: the economy of judgment operates in the production of AI-mediated science with the same structural logic identified in AI-mediated societies, and the resulting redistribution of verification costs can be observed in a fully documented public corpus.

**Table 1. Public case corpus and current maturity labels**

| Document  | Core role in the present article | Public maturity used here                         |
|-----------|----------------------------------|---|
| Paper I   | Physics-branch corpus item       | Theorem-level local PDE result                    |
| Paper II  | Physics-branch corpus item       | Theorem-and-protocol reduction paper              |
| Paper III | Physics-branch corpus item       | Theorem-level branch construction                 |
| Paper IV  | Physics-branch corpus item       | Executed benchmark with explicit scope boundary   |
| Paper V   | Physics-branch corpus item       | Executed negative discriminator on a derived test |
| Paper A   | Companion manuscript             | Conditional comparative novelty assessment        |
| Paper B   | Companion manuscript             | Analytic case study of the human-AI collaboration |

The interpretive rule used throughout is strict: Papers I–V are treated as the primary public case object; Paper A and Paper B are cited as companion manuscripts of the same project, not as external validation.

## 5. Analysis along the six axes

### 5.1 Time of interpretation

The LLM compresses the time required to produce derivations, proofs, textual drafts, audit structures, and formal consistency checks. Paper B’s functional attribution matrix assigns derivational expansion and textual synthesis as AI-dominant tasks across all five physics papers. What would have taken a trained physicist weeks of derivational work was produced in days.

But the compression of production time does not compress the time required to judge whether the product is physically meaningful. A referee evaluating Paper I must verify whether the eigenbasis of the principal symbol is genuinely complete, whether the constraint propagation argument is watertight, and whether the patched admissibility conditions are physically sensible — not merely formally consistent. A referee evaluating Paper IV must assess whether the structural coefficient  $\xi$  carries physical significance or is merely a numerical coincidence. None of these judgments can be accelerated by the same tools that accelerated the production.

The mechanism is precisely P1 of the economy of judgment: when algorithmic mediation compresses production time without making provenance publicly reconstructible, verification costs shift downstream. In scientific production, “downstream” means the disciplinary community — referees, seminar audiences, independent research groups, and editorial boards.

The generalization is immediate. In any ecosystem where AI lowers the barriers to scientific production, the cost of verification becomes the structural bottleneck. The volume of preprints on arXiv, bioRxiv, and Zenodo can increase without bound; the referee pool, the seminar time, and the independent replication capacity cannot increase at the same rate. The economy of judgment predicts that this asymmetry will generate increasing strain on verification infrastructure — a prediction that is already observable in the rising difficulty of securing timely peer review across multiple disciplines.

## 5.2 Visibility

Zenodo makes each paper visible and citable with a DOI within hours of deposit. This is a genuine democratization of visibility: any researcher, regardless of institutional affiliation, can make their work publicly accessible and permanently archived. The seven GDE papers are currently findable, citable, and downloadable by anyone.

But visibility is not credibility. A preprint deposited by an independent non-physicist does not carry the same institutional weight as a paper submitted by an established research group through a recognized journal. The DOI guarantees discoverability; it does not guarantee that anyone in the relevant disciplinary community will invest the time to read, evaluate, or contest the work. Visibility is democratized; interpretive authority is not.

This maps directly onto the second axis of the economy of judgment. The mechanism is P2: increases in visibility without activated contestability tend to concentrate interpretive authority in the hands of producers rather than improve the quality of collective judgment. In the present case, the seven papers are maximally visible (public, DOI-equipped, open access) but minimally contested (no peer review, no independent verification, no seminar discussion as of the deposit date). The interpretive authority therefore rests entirely with the producer — which is precisely the regime that the economy of judgment identifies as epistemically fragile.

The generalization concerns the growing tension between open-access infrastructure and verification infrastructure. Preprint servers have massively increased visibility; they have not proportionally increased the institutional capacity for verification. The economy of judgment suggests that this gap is not a temporary inefficiency but a structural feature of an ecosystem in which the costs of production and the costs of visibility have been reduced while the costs of verification remain unchanged or increase.

## 5.3 Provenance

Paper B attempts to reconstruct who produced what through a functional attribution matrix. The matrix assigns each task (research question, formalism choice, equation derivation, theorem work, benchmark choice, numerical pipeline, scope boundary, text drafting, quality-gate discipline) to one of three categories: human-dominant (H), AI-dominant (A), or co-constructed (HA). The result is informative but self-reported, and the paper explicitly acknowledges that it cannot be treated as externally validated.

Crucially, the granular evidence that would make provenance fully verifiable — session logs, model-specific task ledgers, prompt histories, discarded branches, counterfactual arms — is absent from the current evidence package. Paper B declares this absence explicitly rather than concealing it. The provenance is therefore described but not documented — a condition structurally analogous to the ChatAmsterdam case in the economy-of-judgment paper, where the provider and model are declared in the algorithm register but the source registration links to an internal repository inaccessible from outside.

This is a new problem for scientific authorship. Traditional authorship norms assume that the named author can account for the intellectual content of the work. When the intellectual content is co-



produced with an LLM, the author can account for the architectural decisions (which questions to ask, which scope boundaries to impose, which hard nulls to declare) but cannot independently verify the derivational content (whether the block-triangular structure of the principal symbol is correct, whether the hedgehog scaling argument is valid, whether the constraint propagation proof is complete). The provenance of the intellectual content is distributed across a human-machine chain in which neither participant has full epistemic access to the other's contribution.

The generalization is that AI-assisted science generates a new category of provenance problem. It is not the traditional problem of plagiarism (claiming credit for someone else's work) or ghost authorship (concealing a contributor). It is the problem of opaque co-production: the final output integrates contributions from sources with fundamentally different epistemic characteristics (human architectural judgment and machine derivational throughput), and no existing standard of scientific authorship adequately tracks this integration.

### 5.4 Capability: the distributed capability deficit

This is the axis where the case exposes a genuinely new phenomenon. In the present case, the human producer cannot independently verify the correctness of tensor derivations, constraint propagation arguments, or RG closure claims. The LLM cannot independently verify the physical significance of the constructions — whether the constitutive law captures real superfluid physics, whether the hedgehog defect corresponds to any known particle, whether the deep invariant has empirical content beyond a numerical coincidence.

Neither participant in the production chain possesses the complete capability required to certify the product. This is what I call a distributed capability deficit: a regime in which the production of formally coherent scientific output outruns the verification capacity of every participant in the production chain.

The distributed capability deficit is the structural mechanism behind sophisticated wrongness (Viliotti 2026c): the production of formally coherent, notionally well-structured scientific text that is nonetheless physically empty, subtly unsound, or non-falsifiable in ways that a non-specialist cannot reliably detect and that the generating system cannot self-diagnose. Sophisticated wrongness is the product; distributed capability deficit is the mechanism.

The implications are specific and testable. If the capability deficit is distributed, then verification must come from outside the production chain. In science, this means the disciplinary community: referees, seminar audiences, independent research groups. But these external verifiers face their own economy of judgment: they must decide whether to invest scarce time in evaluating work whose provenance is partially opaque and whose producer lacks field credentials. The capability deficit inside the production chain thus generates a verification demand outside it — but whether that demand is met depends on whether the external community has sufficient resources and incentives to respond.

This connects directly to P3 of the economy of judgment: deficits of capability amplify the effects of provenance opacity, because they make independent verification more costly. In the present case, the opacity of the H/A/HA attribution (axis 3) is amplified by the distributed capability deficit (axis 4): even if the provenance were fully documented at session level, a referee would still need to verify whether the derivations carry physical content — and the producer cannot help with that verification because the producer does not possess the field capability to do so.

The generalization extends beyond non-physicists using LLMs. Even a professional physicist using an LLM for derivational assistance faces a version of the same deficit: the LLM may produce a formally valid derivation that happens to be physically vacuous (correct algebra, wrong physics), and the physicist must invest the verification effort that would have been built into the manual derivation process. The distributed capability deficit is therefore not a property of non-expert producers alone; it

is a structural feature of any production chain in which formal derivation is delegated to a system that does not understand the physical semantics of its output. What changes between the non-physicist case and the physicist case is the magnitude of the deficit, not its structure. Collins and Evans's (2002) distinction between contributory and interactional expertise is useful here: the human actor in this case may possess interactional expertise sufficient to direct a programme, without possessing the contributory expertise required to certify its formal content independently. Humphreys (2004) makes a complementary point about computational science more broadly: when scientific work is extended by computational systems, the locus of epistemic responsibility shifts in ways that traditional accounts of scientific authorship do not capture.

## 5.5 Contestability

The GDE corpus is unusually rich in offered contestability. Paper A defines six hard nulls: HN-1 (field-redefinition reducibility), HN-2 (Einstein-aether collapse), HN-3 (superfluid dark matter collapse), HN-4 (cosmology-linked  $\chi^2$  forced outside range), HN-5 (Jacobson-Verlinde empirical indistinguishability), HN-6 (no empirical separation from neighbors). Paper V defines four protocol hard nulls: N1 (median acceleration scale outside window), N2 (residual RMS too high), N3 (no significant correlation — activated), N4 (external field dominates while internal does not). Paper B defines three collaboration hard nulls: HN-B1 (single-prompt reproduction), HN-B2 (substantial errors found by specialists), HN-B3 (no expert engagement within a reasonable horizon).

This is contestability designed into the product. The producer has pre-declared the conditions under which the claims would fail. This is epistemically valuable and, in the landscape of AI-assisted scientific production, unusual.

But offered contestability is not activated contestability. A hard null is operationally empty until someone invests the cost of testing it. HN-2, for example, requires a specialist in Einstein-aether theory to check whether the patched GDE action can be rewritten as an aether action with spectators. HN-B2 requires an independent physicist to evaluate whether the mathematics is sound. As of the deposit date, none of these hard nulls has been activated by an external agent.

This exposes a fundamental asymmetry in the economy of scientific judgment. The producer bears the cost of designing contestability; the community bears the cost of activating it. If the community does not invest — because the work lacks institutional signaling, because the producer lacks credentials, because the opportunity cost of reading seven papers by an unknown non-physicist is too high — then the hard nulls remain formally present but operationally inert. Contestability without activation is transparency without verification.

The generalization is that contestability in science is a habitat-dependent good. It requires not only that the producer expose failure modes but that there exist agents with the capability, time, and incentive to test them. The economy of judgment predicts that, as AI lowers the cost of production and increases the volume of scientifically formatted output, the cost of activated contestability will rise — because the pool of competent verifiers does not scale with the volume of production.

## 5.6 Institutional pluralism

The GDE case reveals a specific form of institutional concentration. The production chain depends on two commercial LLMs (ChatGPT Pro 5.4 and Claude Opus 4.6) as the primary channels of formal translation. The archival infrastructure is a single platform (Zenodo). The disciplinary community that could verify the work (gravitational physics, modified gravity, mathematical relativity) has not yet engaged.

This is a contraction of the mediators of scientific judgment. In a traditional production chain, a physics paper passes through multiple institutional filters: department seminars, conference

presentations, journal referees, editorial boards, citation networks, replication studies. Each filter represents a distinct site of judgment with its own criteria, competences, and incentive structures. The institutional pluralism of this chain is what makes scientific knowledge robust — not because any single filter is infallible, but because errors that survive one filter are likely to be caught by another.

In the present case, the production chain bypasses most of these filters. The work goes directly from the human-LLM interface to a public archive, without passing through seminars, conferences, departmental scrutiny, or journal review. The only institutional filter activated so far is the self-administered audit structure of the programme itself (hard nulls, maturity labels, scope boundaries). Self-administered contestability is valuable but is not a substitute for pluralistic institutional verification.

The generalization is that AI-mediated science, if it remains outside traditional institutional channels, risks creating a parallel production ecosystem in which volume is high, visibility is democratized, but the institutional pluralism that enables collective verification is absent. The economy of judgment predicts that this regime will not be self-correcting: the producers will continue to produce (because the cost is low), the archives will continue to store (because that is their function), but the verification gap will widen unless institutional mechanisms are created or strengthened to bridge it.

## 6. The three-layer mechanism in AI-mediated science

The analysis of the six axes reveals how the three layers of the economy of judgment operate in scientific production.

In the allocation layer, LLM mediation compresses production costs (derivational throughput, textual synthesis, formal consistency checking) and democratizes visibility (open-access deposition with DOI). The result is a dramatic increase in the volume of scientifically formatted output at decreasing marginal cost. This is not, by itself, a problem — it is an expansion of the production frontier.

In the epistemic traceability layer, the same mediation generates provenance opacity (the H/A/HA boundary is declared but not documented) and a distributed capability deficit (neither the human nor the LLM can certify the product independently). The result is that the formal coherence of the output outpaces the epistemic traceability of its production — which is the structural mechanism of sophisticated wrongness.

In the institutional correction layer, the producer offers contestability (hard nulls, maturity labels, scope boundaries) but cannot activate it unilaterally. Activation requires external agents with competence, time, and incentive. Meanwhile, the institutional pluralism of the verification chain is contracted: the work bypasses seminars, conferences, journal review, and departmental scrutiny.

The combined effect is a regime in which production costs decrease, verification costs are shifted downstream, and the institutional capacity to absorb those costs does not increase. This is not a failure of the individual case — the GDE programme is unusually disciplined in exposing its own limits. It is a structural feature of any production chain in which AI mediation lowers the cost of generating scientifically formatted output without proportionally increasing the institutional capacity for verification.

## 7. Four middle-range propositions for AI-mediated science

The following propositions reformulate the general economy-of-judgment propositions (Viliotti 2026b) for the scientific production domain. They are designed to be testable against independent cases.

**P1-sci.** When LLM mediation compresses scientific production time without making epistemic provenance reconstructible, the costs of verification shift to the disciplinary community — referees, editorial boards, seminar audiences, and independent research groups.

**P2-sci.** Increases in visibility through open-access deposition, in the absence of activated contestability (peer review, independent verification, replication), concentrate interpretive authority in the hands of producers rather than improving the quality of collective scientific judgment.

**P3-sci.** The distributed capability deficit (the human cannot verify the formal content; the LLM cannot verify the physical content) amplifies the effects of provenance opacity, because no participant in the production chain can independently certify the product — making external verification both more necessary and more costly.

**P4-sci.** Institutional pluralism in science (journals, seminars, conferences, independent groups, replication studies) moderates the concentration of interpretive authority only when effective resources exist for verification, replication, and disciplinary memory. Without such resources, the formal availability of multiple institutional channels does not translate into effective collective verification.

These propositions do not exhaust the theory; they operationalize it for the scientific domain. Each is designed to generate observable implications in cases beyond the one studied here.

## 8. Falsifiability architecture: hard nulls, boundary conditions, and portability tests

A central requirement of this manuscript is that it not remain rhetorically insulated. The argument is therefore bound to an explicit falsifiability architecture. That architecture has three layers: first, the four middle-range propositions formulated for AI-mediated science; second, a set of hard nulls directed at the paper’s own framework claim; third, boundary conditions and portability tests that specify where the argument should travel and where it should stop.

The framework-level hard nulls are narrow and operational. HN-P8-1 states that if the six axes and three layers add no diagnostic value beyond adjacent frameworks when applied to AI-mediated science, then the framework claim is redundant. HN-P8-2 states that if independent researchers do not find propositions P1-sci through P4-sci useful when applied to cases beyond the one studied here, then the portability claim fails. HN-P8-3 states that if the distributed capability deficit turns out to be merely an artifact of this specific case rather than a structural feature of AI-mediated production chains more broadly, then the generalization claim is too strong.

The case also imports companion falsifiability burdens that cannot be silently absorbed into the present paper. From Paper A, the comparative novelty claim remains conditional because the strongest collapse burden—field-redefinition reducibility—remains open. From Paper B, the collaboration claim remains conditional because independent single-prompt reproduction, specialist identification of substantial physical or mathematical error, or sustained expert non-engagement would all weaken or falsify stronger readings of the case. From Paper V, one derived discriminator has already returned a negative result: protocol hard null N3 was activated, while N1, N2, and N4 were not. This matters because the case is not an unbroken success narrative. Its public record already contains a failed or non-discriminating test.

| Object under test                                      | Negative discriminator or hard null                                 | Current public status | What would count as failure   |
|--|---|-----------------------|---|
| Framework claim of diagnostic surplus                  | HN-P8-1   | Open                  | Independent application shows that adjacent frames explain the same case equally well without the three-layer/six-axis architecture |
| Portability beyond the present case                    | HN-P8-2   | Open                  | Independent researchers do not find P1-sci–P4-sci useful or discriminating on other cases   |
| Distributed capability deficit as structural mechanism | HN-P8-3   | Open                  | Independent cases with strong AI mediation show no verification asymmetry beyond ordinary scholarly labor                           |
| Conditional novelty status of the physics branch       | Paper A collapse nulls (especially field-redefinition reducibility) | Open and conditional  | A neighboring programme absorbs the branch at action or prediction level without residue  |
| Human–AI collaboration claim                           | Paper B HN-B1–HN-B3   | Open and conditional  | Single-prompt reproduction, specialist error finding, or durable expert non-engagement undermines the stronger interpretation       |
| Derived empirical discriminator in the branch          | Paper V N1–N4   | Partly executed       | N3 already activated; the run was non-discriminating rather than confirmatory   |

The paper’s boundary conditions are equally important. First, this article concerns AI-mediated scientific production chains in which large language models participate in derivational expansion, textual synthesis, or formal packaging. It is not a general sociology of all computational science. Second, the article is about the production and verification ecology of claims, not about whether the claims are physically true. Third, the provenance class of the case is analytic, not forensic. Any attempt to turn the paper into a session-level historical reconstruction exceeds the available evidence. Fourth, the framework is interpretive rather than metrical. The paper does not offer a validated index, causal estimation strategy, or predictive scoring system.

These boundary conditions matter because they define what would and would not count as a relevant external test. A useful portability test is not “can this framework explain everything about AI and science?” That is too broad to fail clearly. A useful portability test is whether the same three-layer and six-axis structure helps distinguish cases that would otherwise look superficially similar. For example: a case produced by a domain expert using an LLM with robust session logging should differ from the present case above all on the provenance and capability axes; a laboratory case with heavy instrumentation but weak public archiving should shift the stress point from derivational opacity to data and workflow provenance; a journal-reviewed AI-assisted article should differ from the present case mainly on institutional pluralism and activated contestability.

The most important portability tests are therefore comparative rather than universal. At minimum, four such tests follow. First, apply the codebook to an AI-assisted case in which the human producer is a recognized domain expert and check whether the distributed capability deficit narrows in magnitude without disappearing in structure. Second, apply it to a case with frozen interaction logs and model-

specific ledgers to test whether stronger provenance actually lowers the downstream verification burden. Third, apply it to a case that has already passed peer review and compare whether activated contestability changes the allocation of verification cost. Fourth, apply it outside theoretical science—for example to computational social science or biomedicine—to test whether the same axes remain analytically useful when the relevant verifiers and archives differ.

This architecture also clarifies the paper’s discipline. The manuscript does not ask to be believed because it is reflexive, ethically cautious, or normatively attractive. It asks to be tested. If the framework travels poorly, it fails. If the propositions do not illuminate independent cases, they fail. If the present case turns out to be idiosyncratic in exactly the way critics would predict, the claim must be narrowed accordingly. That is what makes the article falsifiable rather than merely evocative.

## 9. Discussion: governance, authorship norms, and knowledge infrastructure

The analysis suggests that the response to AI-mediated science cannot be limited to optimism about productivity or alarm about misuse. Different problems sit in different layers of the mechanism, and different interventions therefore address different parts of the verification bottleneck. The practical value of the economy-of-judgment framework is that it shows where the bottleneck accumulates, how it is redistributed, and why disclosure by itself does not exhaust the governance problem.

The analysis suggests that the response to AI-mediated science cannot be limited to regulating models or requiring disclosure. Different problems require different interventions, and the three-layer structure of the economy of judgment provides a map.

### 9.1 Addressing the allocation layer

The allocation problem is not that AI produces too much but that the volume of production is decoupled from the capacity for verification. Institutional responses at this layer would include: funding peer review as remunerated labor (recognizing verification as a cost, not a free externality); developing structured triage systems for preprint servers that distinguish different maturity levels; and creating incentive structures that reward verification, replication, and critique alongside original production.

### 9.2 Addressing the epistemic traceability layer

The traceability problem requires new standards for provenance in AI-assisted research. Mandatory disclosure of AI use (as currently required by *Nature*, *Science*, and other journals) addresses one dimension but does not address the deeper problem of opaque co-production. Responses at this layer would include: development of standardized session logging for AI-assisted research; creation of attribution frameworks that distinguish architectural contributions (human) from derivational contributions (AI) from verification contributions (human or external); and integration of provenance standards from the cultural-heritage sector (metadata, documentary chains, archival protocols) into scientific archiving.

### 9.3 Addressing the institutional correction layer

The correction problem is that contestability offered by the producer cannot substitute for contestability activated by the community. Responses at this layer would include: institutional recognition of hard-null design and negative-result publication as valuable scientific contributions; creation of dedicated channels for the verification of AI-assisted work (not to segregate it, but to ensure that the specific verification challenges it poses receive adequate attention); and diversification of the



mediators of scientific judgment to prevent dependence on a small number of LLM providers as the dominant formal-translation channel.

## 9.4 Connecting scientific and cultural infrastructure

A final implication connects this analysis to the broader economy-of-judgment framework. The provenance problems, archival challenges, and institutional-pluralism questions identified here for scientific production are structurally the same problems identified for cultural heritage, libraries, archives, and museums in the prior work (Viliotti 2026a, 2026b). As Herzog (2023) argues in her account of citizen knowledge, the infrastructure of democracy depends on institutions that make knowledge publicly accessible, contestable, and correctable — markets and experts alone do not suffice. Scientific repositories (Zenodo, arXiv, bioRxiv), cultural-heritage data spaces (Europeana, the Common European Data Space for Cultural Heritage), and public-sector algorithm registers face the same structural question: how to maintain reconstructible, contestable, and pluralistically mediated documentary chains in environments where AI compresses production, obscures provenance, and concentrates interpretive authority.

This convergence suggests that governance responses should be designed across sectors rather than in silos. Provenance standards for scientific archives and provenance standards for cultural-heritage data need not be identical, but they face the same structural problem and could benefit from shared diagnostic infrastructure.

Current authorship and disclosure norms address only part of the structure described here. Journal and editorial guidance increasingly requires disclosure of AI-assisted writing and rejects listing AI systems as authors, while leaving responsibility with the named human author (ICMJE, 2025; Thorp, 2023). These norms are necessary. They preserve accountable authorship, prevent the fiction that a model is a responsible scholar, and force some degree of transparency about assistance.

But they solve only the naming problem. They do not solve the provenance problem of opaque co-production. A disclosure that “AI was used” does not reconstruct which parts of the reasoning were architected by the human, which were derivationally expanded by the model, which were co-constructed through iterative prompting, and which were later filtered out. Nor does it solve the capability problem: a human can remain fully responsible in a normative sense while still lacking the contributory expertise needed to certify the formal content independently. This is why current policy guidance is better understood as a floor than as a complete solution.

The present case makes that limit visible. Its disclosure is unusually explicit by current standards, yet the evidentiary boundary remains analytic rather than forensic. The implication is not that disclosure is unhelpful. It is that provenance standards for AI-assisted science will have to become more documentary and less merely declarative if they are to affect verification costs materially. The relevant question is no longer only whether AI use was disclosed, but whether the documentary chain of production is reconstructible enough for outside communities to judge what happened without reverse-engineering the whole process from a polished artifact.

## 10. Limits and research agenda

### 10.1 Strong objections

The first objection is that the paper is self-referential to the point of circularity: the author applies their own framework to their own case. This is true and cannot be eliminated. It can only be managed — through explicit hard nulls, boundary conditions, and the design of propositions that are testable on

independent cases. The value of the paper does not depend on the physics being correct or the collaboration being optimal; it depends on the diagnostic framework being portable.

The second objection is that the case is unique and therefore not generalizable. This is partly true. No single case establishes a general theory. But the propositions P1-sci through P4-sci are formulated independently of the case: they generate predictions that can be tested on any instance of AI-mediated scientific production. The case provides the initial articulation; generalizability depends on future application.

The third objection is that the economy of judgment is merely a relabeling of well-known problems (the replication crisis, the peer-review burden, the open-access paradox). The response is that the framework does not claim to discover new problems. It claims to provide a diagnostic structure that connects problems currently treated in separate literatures: the attention economy, the epistemology of testimony, the governance of platforms, the sociology of scientific knowledge, and the policy of open access. The surplus is in the connection, not in the novelty of any single component.

## 10.2 Limits

The paper does not offer causal proof, quantitative measurement, or validated metrics. It offers a conceptual framework applied to a single documented case. The provenance of the case itself is only partially documented. The physics underlying the case has not been externally validated. The self-referential structure of the analysis introduces a conflict of interest that is declared but not removed.

## 10.3 Research agenda

Three directions follow from the analysis. First, application of the six-axis codebook to independent cases of AI-assisted science — ideally cases in which the human producer is a domain expert, to test whether the distributed capability deficit persists (as predicted) or disappears (as the case-specific objection would suggest). Second, longitudinal study of how peer-review systems respond to AI-assisted submissions, using the three-layer structure to track whether verification costs increase, shift, or are absorbed. Third, comparative analysis between provenance standards in scientific archives and provenance standards in cultural-heritage data spaces, to test whether the convergence hypothesized in section 7.4 is real or merely analogical.

## 11. Conclusion

This paper has argued that AI-mediated scientific production is an instance of the economy of judgment: the same mechanism that redistributes interpretive authority and verification costs across AI-mediated societies operates in the production, verification, contestation, and archiving of scientific knowledge. The six diagnostic axes — time of interpretation, visibility, provenance, capability, contestability, and institutional pluralism — apply to the scientific production chain without ad hoc modification. The three layers — allocation, epistemic traceability, and institutional correction — identify a structural regime in which production costs decrease while verification costs shift downstream onto a disciplinary community whose capacity to absorb them does not increase at the same rate.

The case studied here — a seven-paper emergent-gravity programme produced by a non-physicist with LLM support — is extreme but analytically useful. It makes the mechanism maximally visible because the capability asymmetry is maximal, the provenance is partially opaque, the contestability is offered but not activated, and the institutional pluralism of the verification chain is contracted. The case does not prove that AI-assisted science is doomed or that it should be prohibited. It shows that the conditions under which scientific judgment operates are being redistributed by the same forces that

redistribute judgment in every other AI-mediated environment — and that understanding this redistribution requires an analytical framework, not just optimism or alarm.

The answer to the question in the title — who pays the cost of verification? — is the disciplinary community, the archival infrastructure, and ultimately the public that depends on scientific knowledge being reliable. Protecting the quality of scientific judgment requires investing in the institutions that make verification possible: peer review as funded labor, provenance standards for AI-assisted work, incentive structures for contestation and replication, and a diversity of mediating institutions that prevents the concentration of interpretive authority in a small number of production channels.

The economy of judgment does not prescribe which investments to make. It provides the diagnostic framework for understanding why they are needed, where the costs accumulate, and who is currently bearing them without being compensated. That is the minimum analytical infrastructure required before any governance response can be designed intelligently.

## Author note

The working byline follows the frozen project contract: **Andrea Viliotti and Framework GDE**. Andrea Viliotti is the sole accountable human author and guarantor. **Framework GDE** designates the structured workflow and public project label through which the manuscript was developed; it does not designate an independent human scholar. External journal submission may require venue-specific normalization to a human-only byline, with Framework GDE moved to the author note and contributorship statements. Correspondence details, ORCID identifiers, and any formal affiliation data can therefore be normalized at venue stage without altering the substantive authorship accountability stated here.

## AI-assistance disclosure

This manuscript was produced with declared support from commercial frontier large language models used as project-level tools for formal translation, derivational expansion, local coherence checking, and textual synthesis. The public project record names ChatGPT Pro 5.4 and Claude Opus 4.6 at project level. No phase-by-phase model allocation ledger is available in the present evidence package, and none is claimed here. Human responsibility for conceptualization, scope definition, method, argument, source selection, falsifiability design, and final editorial accountability remains with Andrea Viliotti.

## Conflict of interest

The author applies their own analytical framework to a case closely connected to their own public scientific production. This creates a self-referential structure and a conflict of interest that is declared, managed through explicit hard nulls, companion-manuscript disclosure, analytic-not-forensic evidence classification, and open scope boundaries, but not removed.

## Contributorship (CRediT-style)

|  |  |           |
|--|--|-----------|
| <b>Conceptualization:</b>                        | Andrea   | Viliotti. |
| <b>Methodology:</b>                              | Andrea   | Viliotti. |
| <b>Investigation:</b>                            | Andrea   | Viliotti. |
| <b>Formal analysis:</b>                          | Andrea Viliotti, with declared digital workflow support.   |           |
| <b>Writing – original draft:</b>                 | Andrea Viliotti, with declared digital workflow support.   |           |
| <b>Writing – review &amp; editing:</b>           | Andrea   | Viliotti. |
| <b>Project administration / workflow design:</b> | Andrea Viliotti and Framework GDE (workflow/public label). |           |

Framework GDE is listed here as the project workflow label relevant to process disclosure, not as an accountable human co-author under standard journal criteria.

## Data and materials availability

The case studied in this manuscript is based on public artifacts. The directly analyzed public case materials are the five Zenodo papers of the GDE emergent-gravity branch (Papers I–V) and the two companion manuscripts (Paper A and Paper B) cited in the references. No new private dataset is deposited with this manuscript. Session-level interaction logs, model-specific task ledgers, discarded branches, and counterfactual arms were not available in the present evidence package and are therefore not claimed. The conceptual foundation used in the analytical framework is available in the cited prior works on cognitive rights as habitat rights and the economy of judgment. Freeze of sources and citations for the present manuscript: 2 April 2026 (Europe/Rome).

## References

- Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, 5, 277–280. <https://doi.org/10.1038/s42254-023-00581-4>
- Boiko, D. A., MacKnight, R., Gomes, G., & Li, G. (2023). Autonomous chemical research with large language models. *Nature*, 624, 570–578. <https://doi.org/10.1038/s41586-023-06792-0>
- Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, 28, 151–155. <https://doi.org/10.1087/20150211>
- Bubeck, S., et al. (2025). *Early science acceleration experiments with GPT-5*. arXiv:2511.16072.
- Cohen, T., & Suzor, N. P. (2024). Contesting the public interest in AI governance. *Internet Policy Review*, 13(3). <https://doi.org/10.14763/2024.3.1794>
- Collins, H. M., & Evans, R. (2002). The third wave of science studies: Studies of expertise and experience. *Social Studies of Science*, 32, 235–296.
- Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., & Calvert, S. (2013). *Knowledge infrastructures: Intellectual frameworks and research challenges*. Deep Blue. <https://hdl.handle.net/2027.42/97552>
- Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.
- Goldberg, A., Ullah, I., Khuong, T. G. H., et al. (2024). *Usefulness of LLMs as an author checklist assistant for scientific papers: NeurIPS'24 experiment*. arXiv:2411.03417.
- Herzog, L. (2023). *Citizen knowledge: Markets, experts, and the infrastructure of democracy*. Oxford University Press.
- Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.
- Hyzen, A., Van den Bulck, H., Puppis, M., Kulig, M., & Paulussen, S. (2026). Epistemic welfare and algorithmic recommender systems: Overcoming the epistemic crisis in the digitalized public sphere. *Communication Theory*, 36(1), 46–57. <https://doi.org/10.1093/ct/ctaf018>
- International Committee of Medical Journal Editors. (2025). *Recommendations: AI-assisted technologies in scientific writing*.
- Jaillant, L., Mitchell, O., Ewosh-Opu, E., & Hidalgo Urbaneja, M. (2025). How can we improve the diversity of archival collections with AI? Opportunities, risks, and solutions. *AI & Society*, 40(6), 4447–4459. <https://doi.org/10.1007/s00146-025-02222-z>
- Liang, W., et al. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 1(8). <https://doi.org/10.1056/AIoa2400196>
- Lu, C., et al. (2024). *The AI Scientist: Towards fully automated open-ended scientific discovery*. arXiv:2408.06292.
- Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627, 49–58. <https://doi.org/10.1038/s41586-024-07146-0>
- Romera-Paredes, B., Barekatin, M., Novikov, A., et al. (2024). Mathematical discoveries from program search with large language models. *Nature*, 625, 468–475. <https://doi.org/10.1038/s41586-023-06924-6>
- Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, communications, and the public interest* (pp. 37–72). Johns Hopkins Press.

- Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature*, 614, 214–216. <https://doi.org/10.1038/d41586-023-00340-6>
- Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, 379, 313. <https://doi.org/10.1126/science.adg7879>
- Viliotti, A. (2026a). *Diritti cognitivi come diritti di habitat nell'era dell'IA: Scuole, archivi, biblioteche, musei e territori come infrastrutture delle comunità educanti*. Zenodo. <https://doi.org/10.5281/zenodo.18925752>
- Viliotti, A. (2026b). *L'economia del giudizio nelle società mediate dall'IA: Diritti cognitivi come diritti di habitat tra autorità interpretativa, costi della verifica e istituzioni del giudizio*. Zenodo. <https://doi.org/10.5281/zenodo.19164007>
- Viliotti, A. (2026c). *Local strong hyperbolicity in a patched single-medium effective theory for emergent gravity*. Zenodo. <https://doi.org/10.5281/zenodo.19319073>
- Viliotti, A. (2026d). *The Einstein window in a patched single-medium effective theory for emergent gravity: Reduction theorem, gravitational-wave corollary, and confirmatory protocol*. Zenodo. <https://doi.org/10.5281/zenodo.19320413>
- Viliotti, A. (2026e). *From a local quantum medium to a patched effective theory: RG closure and an explicit hedgehog defect in emergent gravity*. Zenodo. <https://doi.org/10.5281/zenodo.19320527>
- Viliotti, A., & Framework GDE. (2026f). *The deep acceleration scale in a single-medium emergent gravity:  $\xi = 1.38 \pm 0.11$  from SPARC, Planck, and cross-habitat constraints*. Zenodo. <https://doi.org/10.5281/zenodo.19348070>
- Viliotti, A. (2026g). *Per-galaxy RAR residuals in a single-medium emergent gravity: Testing the non-equilibrium prediction against the external field effect on SPARC*. Zenodo. <https://doi.org/10.5281/zenodo.19353740>
- Viliotti, A., & Framework GDE. (2026h). *A single quantum medium as the source of gravity, matter, and cosmic acceleration: A comparative foundations assessment of novelty, competitive positioning, and distinguishable predictions*. Zenodo. <https://doi.org/10.5281/zenodo.19373599>
- Viliotti, A., & Framework GDE. (2026i). *AI-assisted theoretical-physics programme construction by a non-physicist: A falsifiable analytic case study from the GDE emergent-gravity branch*. Zenodo. <https://doi.org/10.5281/zenodo.19373683>

## Appendix A. Six-axis codebook applied to the case

| Axis                    | Layer        | Observable in the case   | Generalization  |
|-------------------------|--------------|--|---|
| Time of interpretation  | Allocation   | LLM compresses derivation from weeks to days; referee verification time unchanged or increased | Production speedup without verification speedup shifts costs downstream                         |
| Visibility              | Allocation   | Zenodo + DOI = immediate global visibility; zero institutional gatekeeping                     | Open-access democratizes visibility without democratizing credibility                           |
| Provenance              | Traceability | H/A/HA matrix declared but not documented at session level; no logs, no counterfactual arms    | AI co-production generates opaque provenance unaddressed by current authorship norms            |
| Capability              | Traceability | Human lacks field expertise; LLM lacks physical judgment; distributed capability deficit       | No participant can certify the output; verification must come from outside the production chain |
| Contestability          | Correction   | 13+ hard nulls pre-declared across the corpus; none activated externally as of deposit         | Offered contestability without activation is transparency without verification                  |
| Institutional pluralism | Correction   | Bypasses seminars, conferences, journal review; depends on 2 LLMs and 1 archive                | Contracted verification chain reduces collective error-correction capacity                      |

## Appendix B. Public case corpus and maturity map

| Document  | Public date | Role in the present paper  | Public maturity status used here                                   |
|---|-------------|--|--|
| Paper I. <i>Local Strong Hyperbolicity in a Patched Single-Medium Effective Theory for Emergent Gravity</i> | 29 Mar 2026 | Part of the public case corpus                                     | Theorem-level local PDE result                                     |
| Paper II. <i>The Einstein Window in a Patched Single-Medium Effective Theory for Emergent Gravity</i>       | 29 Mar 2026 | Part of the public case corpus                                     | Theorem-and-protocol reduction paper                               |
| Paper III. <i>From a Local Quantum Medium to a Patched Effective Theory</i>                                 | 29 Mar 2026 | Part of the public case corpus                                     | Theorem-level branch construction                                  |
| Paper IV. <i>The Deep Acceleration Scale in a Single-Medium Emergent Gravity</i>                            | 30 Mar 2026 | Part of the public case corpus                                     | Executed benchmark with explicit scope boundary                    |
| Paper V. <i>Per-Galaxy RAR Residuals in a Single-Medium Emergent Gravity</i>                                | 31 Mar 2026 | Part of the public case corpus                                     | Executed negative discriminator on a derived test                  |
| Paper A. <i>A Single Quantum Medium as the Source of Gravity, Matter, and Cosmic Acceleration</i>           | 1 Apr 2026  | Companion manuscript for conditional novelty/non-collapse status   | Comparative foundations assessment; conditional novelty verdict    |
| Paper B. <i>AI-Assisted Theoretical-Physics Programme Construction by a Non-Physicist</i>                   | 1 Apr 2026  | Companion manuscript for the collaboration and provenance boundary | Analytic case study; explicit analytic-not-forensic evidence class |

**Interpretive rule used in the manuscript.** Papers I–V constitute the primary public object. Papers A and B are used as companion manuscripts of the same project, not as external validation.