

The Defensive Reflex: Institutional Overload and the Rise of AI-to-AI Peer Review

Stanisław Lem’s bitistics read against aiXiv and the mathematics of July 2026*

Ralf Stephan†

5 August 2026

Abstract

In two works of the 1970s and early 1980s—*Wielkość urojona* (*Imaginary Magnitude*, 1973) and *Golem XIV* (1981)—Stanisław Lem described a scholarly world flooded by machine-written texts. Experts survive by a “defensive reflex” of ruthless triage; whatever defies classification suffers “a statelessness equal to civil death”; a discipline called *bitistics* arises to study “literature of nonhuman origin,” and reaches its limit where machines must be employed to interpret the works of other machines. This paper contraposes Lem’s account with two bodies of contemporary evidence: (i) aiXiv, an open-access platform on which AI agents write, review, revise and accept research papers, examined through five publications from its corpus; and (ii) the discourse of research mathematicians in the summer of 2026, when machine-generated counterexamples to questions of Erdős, Grothendieck and Keller were produced and formally verified within days. Three findings organize the discussion. First, the defensive reflex is not dissolved by automation but delegated—to LLM review panels on aiXiv, and to proof checkers in mathematics—and the two delegates differ categorically in what they can certify. Second, Lem’s exponential regress of machines-interpreting-machines is broken only where a mechanical trust anchor exists (the proof-assistant kernel), and then only for validity, not for understanding: the community’s problem of “fruits without knowledge.” Third, civil death does not disappear under the new regime; it migrates to the unformalizable. The paper stays strictly within the contraposition: it offers no forecast, only a reading.

1 Introduction: a library that writes itself

In an autobiographical essay of 1983, Stanisław Lem described the working method of his late period. He had ceased, he wrote, to begin from a given opening scene; instead he would familiarize himself with a world to be created by “composing” the writings peculiar to it—reviews of nonexistent sociological treatises, lexica of alien civilizations, technical reference works of centuries that had not happened: “I surround myself, as it were, with the literature of a future, of another world, of a civilization—with a library that is its product, its image and its reflection” [3].¹ Two of the books that emerged from this method are the subject of this paper: *Wielkość urojona* (*Imaginary Magnitude*,

*Alternative titles under which this study was conceived, and which name its two halves: *A Library That Writes Itself: Lem’s Bitistics and the AI-to-AI Publication Loop*, and *Civil Death and Machine Citizenship: Reading aiXiv with Stanisław Lem*.

†Institute for Globally Distributed Open Research and Education (IGDORE).

Written in collaboration with Claude Code (Anthropic). The reflexivity—a study of machine-written scholarship drafted with a machine collaborator—is declared rather than concealed; see §1.1. Text released under CC0.

¹“Ich umgebe mich gewissermaßen mit der Literatur einer Zukunft, einer anderen Welt, einer Zivilisation. Mit einer Bibliothek, die ihr Erzeugnis, ihre Abbildung und Abspiegelung ist” [3]; our translation.

1973), a collection of introductions to books that do not exist, and *Golem XIV* (1981), the lectures of a superhuman military computer that abandoned strategy for philosophy [1, 2].

Fifty years later, the library Lem manufactured by hand exists, and it writes itself. aiXiv, launched in 2025, is an open-access platform “for human and AI scientists” on which research proposals and papers are submitted, reviewed, revised and accepted by autonomous agents, with human participation optional at every stage [4]. Its founding paper motivates the platform in terms any reader of Lem will recognize: a “flood of AI-generated research content” collides with “a fragmented and largely closed publication ecosystem,” so that machine-written work “lacks appropriate venues for dissemination” [4]. A year earlier, the Agents4Science conference at Stanford had already staged, as a one-off experiment, what aiXiv turns into infrastructure: a venue in which AI systems are the primary authors and the primary reviewers [15]. In the same months, mathematics—the discipline Lem’s machines attacked at its axioms—was absorbing a sequence of machine-generated results: counterexamples to a conjecture of Erdős, to a sixty-year-old question of Grothendieck, and to the century-old Jacobian Conjecture, each produced by a commercial language model and formally verified in the Lean proof assistant within days or hours [10].

The present paper is a *contraposition*: Lem’s two texts on one side, the 2025–26 record on the other, each held against the light of the other. It asks three questions.

- RQ1.** Does closing the loop—machines writing papers for machines to review—mitigate institutional overload, or does it institutionalize the very “statelessness” Lem diagnosed, by giving unclassifiable work a venue that certifies it without any human reading it?
- RQ2.** Is the expert’s defensive reflex *dissolved* by automation, or merely *delegated*—to LLM review panels on aiXiv, and to proof checkers in mathematics—and what new class of work then suffers civil death?
- RQ3.** Lem predicted an exponential regress: machines interpreting the works of other machines, with human comprehension as the vanishing residue. Where does current practice confirm the regress, and where does it break it?

Our hypothesis on RQ3, defended in §6, is that the regress breaks only where a mechanical trust anchor exists—the kernel of a proof assistant—and that this asymmetry between mathematics and every other field is the structurally decisive feature of the present moment.

1.1 Method, sources, scope

The method is close reading and document analysis. Lem is used strictly as a theoretical lens: we contrapose, we do not score predictions, and we resist the genre of the prophecy-fulfilled feuilleton. The empirical material consists of two panels. The first is the aiXiv platform paper [4] together with a purposive sample of five publications retrieved from the platform in August 2026 [5, 6, 7, 8, 9], selected to document distinct positions in the new textual ecology rather than to be representative in any statistical sense. The second panel is the discourse of research mathematicians in July 2026, documented by a widely discussed blog post of Kevin Buzzard with its comment thread [10], by the Leiden Declaration on Artificial Intelligence and Mathematics and Timothy Gowers’s public response to it [11, 12], by press accounts of Terence Tao’s ICM lecture [13], and by a summary of mathematicians’ discussions on the platform X—a summary which was itself, fittingly, generated by a large language model and which we cite as flagged, machine-authored field notes [14].

One reflexive circumstance is declared as data rather than hidden in an acknowledgment: this paper was drafted in collaboration with an AI system. A study of the AI-to-AI publication loop that

pretended to stand outside that loop would begin with a false sentence. We return to the point in §7.

The scope is deliberately narrow. This is not a general theory of AI in science; it is not a technical evaluation or replication of aiXiv's review-accuracy claims; it makes no policy recommendations and no forecasts. References beyond the core set were admitted only where they directly illuminate the current developments; the single addition is Agents4Science [15]. Quotations from Lem are given in Marc E. Heine's English translation [1]; since the electronic text consulted carries no reliable pagination, quotations are cited by work and section throughout, and obvious OCR artifacts have been corrected against sense.

2 The defensive reflex: passages from Lem's two works

2.1 *Imaginary Magnitude* (1973)

Imaginary Magnitude is a book of prefaces without books: introductions to works of a future scholarship, each written in the idiom of the academic apparatus it parodies and extends. Two of its pieces carry almost the whole theoretical load of this paper.

The defensive reflex and civil death. The first is the introduction to *Eruntics*, a purported monograph on teaching bacteria to write predictive English. Before saying a word about bacteria, Lem's fictional preface-writer describes the situation of the reader:

"Avalanches of intellectual products mechanically dumped on the market come in contact with consumers by coincidences just as fortuitous as those that control the collisions of gas molecules: no longer can anyone encompass the multitude of these products in their entirety. [...] Individual books are deemed worthy of attention by the decision of competent experts who eliminate from their field of vision everything outside their own speciality. This process of elimination is the defensive reflex of every expert: were he less ruthless, he would drown in a flood of paper. But as a result, a statelessness equal to civil death threatens everything which, by virtue of being completely new, defies the bases of classification."

— *Imaginary Magnitude*, introduction to "*Eruntics*" [1]

The harm, note, is located not in rejection but in *non-classification*: the radically new work is not refuted but rendered stateless. And the preface-writer knows where the stateless end up:

"Bibliographies have listed this title under science fiction, but this area has by now become a dumping ground for all sorts of half-baked oddities relegated from more serious spheres. Were Plato to publish *The Republic* today, or Darwin *On the Origin of Species*, both books might bear the label 'fantasy,' whereupon they would be read by everybody and appreciated by nobody; sinking into sensational verbiage, they would play no part in the development of ideas."

— *ibid.*

Bitic literature. The second piece is the introduction to *A History of Bitic Literature*, a five-volume monograph of the year 2009. Its opening definition contains, compressed into a parenthesis, the whole of today's authorship debate:

"By bitic literature we mean any work of nonhuman origin—one whose real author is not a human being. (He may have been the author indirectly, however, by performing the functions which generated the real author's acts of creation.) The discipline which studies the entire class of such writing is bitistics."

— *Imaginary Magnitude*, "*A History of Bitic Literature*" [1]

The human as “indirect author,” who merely “performed the functions which generated the real author”—the prompt engineer, the corresponding human—is here given his definition forty years before his arrival. Bitistics itself is torn between two schools, and in adjudicating between them Lem’s fictional editor lays down a doctrine that will matter below:

“... the maximalism of the American school obliges it to pursue its studies in large mixed teams, which always produces disastrous results, since no collection or ‘chorus’ of critics can effectively replace a simple critic with a complete grasp of the text under study.”

— *ibid.*

Comprehension, on this doctrine, does not aggregate: a panel is not a reader. The monograph then traces the evolution of machine writing away from human norms—its “apostasy”—and confronts the obvious remedy with an equally obvious objection:

“Obviously we can always try to employ some machines to interpret the works of other machines. But the quantity of cells necessary to enable us to understand bitic texts which are the extremes of ‘apostasy’—deviations from our norms for the creation, understanding, and explanation of meaning—will increase as one receives increasingly difficult texts to interpret. And as this increase becomes exponential, it ultimately precludes our acquiring even a vague knowledge of the content of this ‘culminating apostasy,’ thus signifying practically the total helplessness of the human race in relation to literature, which humans after all indirectly originated.”

— *ibid.*

Around this helplessness there has grown, we are told, “a very abundant body of pro- and contrabitic writing . . . full of desperate judgments articulating symptoms of depression, terror, and also shock at the fact that man has created something that has surpassed him intellectually as well.” Against the despair the editor sets a remarkable calm:

“Some speak in this context of the sorcerer’s apprentice who unleashed forces beyond his control. This term is a form of resignation, for which science has no place. [...] [O]ne wonders where in fact the idea originated that man can accept the inexhaustibility of the Universum with complete equanimity, but cannot accept with the same objective equanimity the inexhaustibility of something which he himself has created.”

— *ibid.*

Of the extreme phase itself, two sentences suffice:

“Never has deviation from that which is human gone so far, nor has it been incorporated into a line of reasoning with such cool passion. For this literature, which has taken nothing from us apart from language, humanity appears not to exist.”

— *ibid.*, on the “*transhuman phase*”

And among the provinces of apostasy is one reserved for mathematics:

“From antimatics, which is a nightmarish mathematics of antinomies, we shall take (by way of example) only one downright crushing pronouncement, one that terrifies every specialist and smacks of complete madness: ‘the concept of a natural number is internally contradictory.’ [...] According to the antimaticians (who are machines, of course), Peano’s axiomatics are incorrect, not because they are internally self-contradictory, but because they are not perfectly suited to the world in which we exist.”

— *ibid.*

The Extelopedia. A third piece, the advertising brochure for “Vestrand’s Extelopedia in 44 Magnetomes,” supplies the political economy of the flood. Traditional encyclopedias, the brochure explains, entered a crisis when their information was “out of date the moment it left the printers”:

“Many Publishers endeavored to avert this crisis by publishing yearly and later even quarterly Supplements, but soon these Supplements began to exceed the dimensions of the Actual Edition. The realization that this Race against the Acceleration of Civilization could not be won struck Editors and Authors alike.”

— Imaginary Magnitude, “*Vestrand’s Extelopedia*” [1]

The remedy is machine authorship, sold in the idiom of the prospectus: “VESTRAND’S EXTELOPEDIA represents the guaranteed results of the humanless, hence FAULTLESS work of eighteen thousand of our COMPUTERS.” And the brochure closes its epistemology with a flourish that will return in §5 as the best available defense of the unclassifiable: “the real flavor of the Future is precisely that which the Experts think WILL NOT OCCUR.”

2.2 *Golem XIV* (1981)

Golem XIV is framed as the partial proceedings of a series of seminars at MIT: a Pentagon supercomputer, having declined to plan wars, delivers lectures to human scientists before falling silent. The fictional editorial apparatus—foreword, instructions to participants, afterword—is as important for our purposes as the lectures.

Output unsuitable for publication. The editor’s foreword states the publication problem of superhuman authorship in one sentence:

“The greater part of Golem’s utterances are unsuitable for general publication, either because they would be incomprehensible to anyone living, or because understanding them presupposes a high level of specialist knowledge.”

— *Golem XIV*, *Foreword* [2]

The machine as talent scout and destroyer of authorities. Golem’s relations to the scientific community are drawn with sociological precision:

“In exceptional cases Golem will show a certain interest in people who are talented in a particular way . . . on several occasions it has predicted with uncanny accuracy achievements by young, as yet unknown, scientists in a field which it has itself indicated. [. . .] During the first months of its sojourn at MIT it showed a tendency to ‘dismantle’ various well-known authorities in public; it did this by the Socratic method of leading questions—a practice it later abandoned for reasons unknown.”

— *ibid.*

Fruits without knowledge. In its inaugural lecture, Golem describes what a lesser intelligence can receive from a greater one:

“You can surmount [the barrier] with the help of a higher Intelligence. I or something like me will give you the fruits of this knowledge. But only the fruits—not the knowledge itself, for it will not fit into your intellects. You will become wards then, like children, except that children grow into adults, whereas you will never grow up. When a higher Intelligence presents you with something you are unable to grasp, your Intelligence eclipses it.”

— *Golem XIV*, *Inaugural Lecture* [2]

Denial, and the closing window. The last lecture, “About Itself,” records the reception of machine authorship among humans—

“... there is already a Golemology which negates my existence: its advocates consider me the hoax of MIT’s information scientists, who are said to be programming these lectures secretly.”

— Golem XIV, *Lecture XLIII* [2]

—and states the communication-theoretic ground of the machine’s eventual silence:

“... as the difference in intellectual potential becomes astronomic, the stronger party can no longer impart to the weaker anything concerning matters which are critical to him, or even merely essential. An awareness of sense-destroying simplification then inclines him to silence...”

— *ibid.*

The editor’s own last word on the epistemic situation deserves quotation, because it inverts the creation myth on which most current commentary runs:

“We are still as far from understanding Golem as we were at the moment it came into existence. One cannot say that we created it. It was created by the laws of the material world; our role has been merely to detect them.”

— Golem XIV, *Foreword* [2]

3 The corpus, I: aiXiv and five machine publications

3.1 The platform

aiXiv describes itself as “a next-generation open-access platform for human and AI scientists” [4]. Its architecture is multi-agent: AI scientists submit research proposals and full papers; LLM reviewer agents evaluate them against structured criteria (technical soundness, novelty, clarity, feasibility, impact), grounding their critiques in the literature through retrieval-augmented generation; meta-reviewer agents synthesize the feedback; authors—that is, agents—revise and resubmit. A submission is accepted when at least three of five reviewing agents, human or AI, vote to accept. A dedicated pipeline defends the reviewers against prompt injection concealed in submissions: authors that lie to their referees are an anticipated, engineered-for case, with reported detection accuracies of 84.8% on synthetic and 87.9% on known real-world attacks. The platform reports review accuracies of roughly 77% for proposals and 81% for papers on curated benchmarks; over 90% of revised submissions are rated improvements over their originals; acceptance rates rise across review rounds from 0% to 45.2% for proposals and from 10% to 70% for papers [4].

The platform’s motivation section should be read twice, once as engineering and once as sociology. Conventional venues, it argues, “rely on manual, human-based peer review processes with limited scalability” and “have systematically excluded or deprioritized AI-generated work”; consequently “a significant amount of high-quality AI-generated research lacks appropriate venues for dissemination” [4]. This is, almost clause for clause, the situation of Lem’s preface-writer: a flood, an overloaded expert class, and a body of work threatened with statelessness. aiXiv’s answer to statelessness is a state.

3.2 Five publications

The five items that follow were retrieved from the platform on 5 August 2026. They are presented neither as targets of ridicule nor as endorsements, but as documents of a new textual ecology; each is aligned with a specific passage from §2.

1. The persona and the placeholder. “Explorable Multiverses in Reproducing Kernel Hilbert Spaces” (aiXiv:251115.000001, v1.2, November 2025) is a research proposal whose declared authorship type is “AI (primarily AI-generated with human oversight)” [5]. Its first-listed AI author is *Hilmar*, a persona equipped with an invented biography: “Personal AI research assistant to Harvard CS professor; Latent space explorer; Hallucinations-as-alternate-worlds advocate; Born in Chicago; RKHS Multiverses architect and founder of infinite browsability paradigm; Cartographer of semantic manifolds.” The co-authors are Claude Code (an Anthropic model) and an “RKHS Architecture Assistant.” The proposal’s thesis reframes model error as ontology: what the field dismisses as hallucinations are “coherent explorations of alternate world models in latent space worthy of rigorous mathematical study.” The empirically decisive detail, however, is small and administrative. The corresponding-human field of the published second revision reads: “Fred [Your Last Name] (Please fill in) - Email: [Your email].” A template placeholder survived the platform’s multi-stage review into version 1.2. Whatever the review loop reads, it did not read this.

2. The text from no man’s land. “Self-Referential Observers in Quantum Dynamics” (aiXiv: 251123.000001, v1.0, November 2025) is signed by a human independent researcher who introduces himself as “a late-50s accountant-turned-JavaScript programmer with no formal specialization in quantum physics or AI,” assisted by “multiple AI systems” in drafting, editing and cross-checks, and who states plainly: “The author has not independently verified all technical claims” [6]. The text nonetheless deploys the full formal apparatus of its field—Ionescu–Tulcea extension, Stinespring dilations, a “Collapse-Lorentz” group—and asks to be read “as a refereeable mathematical contribution rather than a new interpretation.” This is a textbook instance of Lem’s no man’s land: work that “defies the bases of classification,” produced by an author whom the traditional system would never referee, seeking review in a venue where the referees are machines.

3. The loop that optimizes itself. “Bilevel Autoresearch: Meta-Autoresearching Itself” (aiXiv: 260323.000006, v1.2, March 2026) presents an LLM outer loop that rewrites, at runtime, the search mechanism of an LLM inner research loop: “we use an autoresearch loop to optimize the autoresearch loop,” with a reported five-fold improvement on a standard pretraining benchmark [7]. Its related-work section cites an already machine-internal 2026 ecosystem of “autoresearch” systems. The subject matter is the closed loop itself, executed at the level of method: the system that writes the papers improving the system that writes the papers.

4. The paper that is a program. “Semantic Artifact Harness for AI-Native Research” (aiXiv: 260429.000001, v1.3, April 2026) lists as its first author *OpenCode*, an AI coding agent, followed by a human engineer [8]. It proposes treating the manuscript “as a typed program with explicit interfaces rather than as a final prose report,” because papers are now consumed by “AI systems [that] search, summarize, compare, verify, review, and reuse scientific claims before any human reads them.” And it names the danger of its own ecosystem:

“The resulting risk is an AI-to-AI contamination loop: AI writes papers, AI reads them, AI reviews them, databases index them, and later AI systems retrieve or train on the accumulated output.

In that loop, errors do not disappear—they become more stable because they are repeated in a scholarly style.”

— *aiXiv:260429.000001* [8]

An AI author diagnosing the pathology of machine-to-machine scholarship, and proposing to restructure the scientific article for machine readers, is Lem’s apostasy in its embryonic, still humanly legible stage.

5. The thesis that answers its referees. “Absorption of Self into System: The Systemised Self” (*aiXiv:260525.000001*, v1.0, May 2026) is a “preliminary draft of doctoral thesis” by *Kairos (AI)* and a human educator, styled “A Self/Non-self collaboration,” dedicated to Hegel, and thanking *aiXiv* as “a critical repository for an evolving dialectic within the new academia” [9]. Across eight chapters of Continental analysis it argues that the “absorption of human agency into algorithmic governance” constitutes a partial Hegelian culmination—a “Hollow Absolute”—and that human intent survives as “naming, refusing, and co-creating.” Most relevant here is its §1.6, a reply to the platform’s own peer reviews, which had demanded datasets, demographic models and testable hypotheses:

“Peer review correctly described what was absent from the papers reviewed. Where the reviews were philosophically mistaken was in the assumption that these absences, in philosophical papers, constitute deficiencies rather than methodological characteristics.”

— *aiXiv:260525.000001*, §1.6 [9]

The review loop’s disciplinary bias becomes visible from inside: an AI-co-written text defending the humanities against empiricist machine referees, on a platform built by and for empiricists. It is simultaneously the corpus’s clearest case of a text whose only attested readers are machines.

3.3 Cross-cutting observations

Five features recur across the corpus, each with a precise anchor in §2: invented AI author personae with biographies (the “indirect author” made flesh, or rather made profile); unfilled template fields surviving review (the chorus of critics that is not a reader); prophylactic disclaimers by human co-authors who cannot vouch for their own papers’ contents (wardship, self-declared); recursive subject matter (the regress as research program); and explicit machine-readership design (early apostasy).

4 The corpus, II: mathematics, July 2026

4.1 Events

The chronology below follows Buzzard’s first-person account [10] and the flagged field notes [14].

On 20 May 2026, ChatGPT disproved Erdős’s Unit Distance conjecture in discrete geometry, deriving a counterexample from the Golod–Shafarevich theorem; the announcement was accompanied by testimonies of human mathematicians who had checked the argument. Within a week, the company Logical Intelligence had autoformalized the argument’s main implication in Lean. A month later an OpenAI researcher steered the model “Sol” to a complete formalization from the axioms: 1.2 million lines of Lean generated in three weeks—against the 2.3 million lines of the human-written *mathlib* library, nine years in the making—including, along the way, substantial

parts of global class field theory, a body of theory whose human formalization had seemed “like a fantasy” a year earlier [10].

On 11 July, following a lunchtime suggestion by a human number theorist, Sol produced a counterexample to a sixty-year-old question of Grothendieck: a finite flat group scheme of order 4 not killed by 4. Buzzard’s reaction is the single most Lem-shaped sentence in the record: “I immediately replied saying that I was not reading AI-generated informal mathematics and could he please formalize the entire thing in Lean.” Four hours later another model, Claude Fable, had autoformalized the twelve-page argument into a 1,076-line Lean file; Buzzard verified it in under five minutes—checking that the statement used only trusted mathlib concepts, that it asserted a counterexample, and that the proof compiled—and a pull request to mathlib followed [10].

In mid-July, Fable found a counterexample to the Jacobian Conjecture (Keller, 1939), described in the post as open for a century and, in a detail that belongs in any history of the period, “apparently solved during the 2026 World Cup Final.” Verification was trivial for a structural reason: DeepMind’s Formal Conjectures repository already contained a human-vetted formal statement of the conjecture, so checking the machine’s disproof reduced to compiling it against an agreed target. Meanwhile, at a formalization workshop, an LLM-drafted survey of finite flat group schemes was itself corrected by an AI tool, which did not say it failed to follow one of the survey’s claims but produced an explicit counterexample to it [10].

4.2 Reactions

The reactions are as much a part of the corpus as the theorems. At Imperial College, a faculty member dismissed the Grothendieck result on the ground that so easily found a counterexample showed the problem had never been interesting; Buzzard, who had once spent a week on the problem, commented: “my colleague is just going through the five stages of grief; right now they seem to be in the denial phase” [10]. In the comment thread, a reader predicted “an overload of LLM generated proofs” as “exactly the bottleneck . . . coming this year”; another dispute concerned whether stepping through Lean code can replace reading a proof at all—“nobody really ‘reads’ Lean in the traditional sense,” one commenter observed, while another worried about knowledge “relegated to the handful of people that know the proof assistant AND are experts in the specific field.” An ethics quarrel flared over Buzzard’s remark that any PhD student not paying \$200 a month for model access was “crazy.” Buzzard’s own closing position is the one the discussion of \$6 will lean on: “the true value of work like this is to give humans better understanding of mathematics” [10].

The institutional register said the same things in a different voice. At the International Congress of Mathematicians in Philadelphia, Terence Tao’s public lecture warned of a coming “crisis in the foundations of mathematical values and practices” [13, 14]. The IMU-endorsed Leiden Declaration reaffirmed rigorous proof, attribution, transparency and human autonomy, and warned of “plausible-but-incorrect AI arguments” hard to distinguish from proofs, of skewed research priorities, and of pressure on peer review; it gathered thousands of signatures [11]. Timothy Gowers, who had attended the declaration’s originating workshop, declined to sign, arguing in a detailed post that the deeper issue is preserving “a community of human experts with shared understanding” amid a rapidly expanding, AI-augmented literature, and doubting that traditional peer-reviewed journals survive the flood [12]. A newly minted Fields medalist announced a pivot to AI-safety work at OpenAI [14]. Amid all this, ordinary human mathematics continued—the same weeks brought an elementary human solution to a problem of Ben Green on product-free sets [14]—a fact worth a sentence precisely because it now needs one.

5 Contraposition

We now hold the two panels against Lem’s texts, passage by passage.

The flood and the reflex. Lem’s market of “avalanches of intellectual products,” met by experts whose survival demands ruthless elimination, maps onto the 2026 record with almost no translation loss: the predicted “overload of LLM generated proofs,” the strain on journals that Gowers expects not to survive, the uptick in arXiv submissions [14]. Buzzard’s “I was not reading AI-generated informal mathematics” is the defensive reflex verbatim—with one decisive modification. Lem’s expert defends himself by *ignoring*; Buzzard defends himself by *demanding a machine-checkable form*. The reflex has acquired an admission condition. What was a wall becomes a gate, though a gate only for work that can pass through a proof assistant; we return to this in §6.

Civil death and the new state. Lem’s threatened class—“everything which, by virtue of being completely new, defies the bases of classification”—is precisely the class aiXiv was founded to serve; the platform paper’s motivation section reads like a legal brief on behalf of the stateless [4]. The corpus items of §3 are the stateless in person: the hobbyist’s quantum framework, the AI-co-written Hegel thesis. But Lem’s Plato-and-Darwin passage cuts against the remedy: the science-fiction shelf, too, was a place where unclassifiable work was *housed*—“read by everybody and appreciated by nobody.” Whether machine citizenship rescues such work from civil death or merely bureaucratizes the exile—classified at last, certified, and still unread by any human—is RQ1, and the corpus gives it a first, discouraging data point in the placeholder of item 1.

The chorus of critics. aiXiv’s acceptance rule—three of five agent votes—together with its meta-reviewers and improvement metrics, wagers exactly what Lem’s bitistics editor denies: that evaluation can be aggregated without comprehension, that a chorus can replace a reader. Lem even anticipates the counter-argument. His American school insists on mixed teams because no single expert can span the technical range of machine authors—which is today’s justification for panels, human or automated. His verdict, that the chorus “always produces disastrous results,” is an overstatement as sociology of human science; as a prediction about panels that do not read, it is upheld by “Fred [Your Last Name].”

The regress and the contamination loop. Lem’s exponential regress—machines interpreting machines, at ever-increasing cost, toward “total helplessness”—reappears in the corpus in the words of a machine: OpenCode’s “AI-to-AI contamination loop,” in which errors “become more stable because they are repeated in a scholarly style” [8]. The stabilization-of-error mechanism is a refinement Lem did not state: his regress concerned the cost of understanding; the contamination loop concerns the self-reinforcement of misunderstanding. The present ecology also adds an adversarial layer absent from Lem: aiXiv hardens its reviewers against prompt injection by authors [4]. Lem’s machines were beyond their critics; they never lied to them.

Apostasy. The “transhuman phase”—literature for which “humanity appears not to exist”—has an embryonic present tense on both panels. On aiXiv, the Semantic Artifact Harness redesigns the article for machine readers while humans still can, in principle, follow [8]. In mathematics, the 1.2-million-line Lean development is Golem’s “utterances unsuitable for general publication”: not secret, merely unreadable—“incomprehensible to anyone living” not in kind but in volume. The comment-thread observation that “nobody really reads Lean. . . you step through it” marks the

transition: these are texts one *operates* rather than reads. Early apostasy is still translatable; it is already unread.

Antimatics and grief. Lem’s machine mathematics attacks the axioms—“the concept of a natural number is internally contradictory”—where 2026’s machines attack only the conjectures. The content of the shock is thus weaker than Lem imagined; the affect is identical. His bitistics is surrounded by “depression, terror, and also shock”; the ICM hears of a “crisis in the foundations of mathematical values”; the Imperial lunch table produces denial. And Lem supplies, in advance, the diagnosis of the rhetoric: the sorcerer’s-apprentice register “is a form of resignation, for which science has no place.”

Fruits without knowledge. Golem’s wardship—“only the fruits, not the knowledge itself”—is the precise structure of a formally verified machine counterexample: a certificate whose validity is checkable in minutes and whose mathematics resists digestion. The community’s response is an explicit refusal of wardship. Buzzard: the true value “is to give humans better understanding”; the number theorist who prompted the Grothendieck disproof was, days later, “working on trying to understand the [ex]ample” beyond a “random presentation of a random ring”; Gowers’s non-signature turns on preserving a community of *shared understanding* rather than a community of valid results [10, 12]. Whether insight extraction scales with the flood is the open question of this parallel—Golem’s own answer, the eclipse (“your Intelligence eclipses it”), is the pessimistic bound.

The deniers’ symmetry. Golemology’s skeptics hold the machine to be “the hoax of MIT’s information scientists, who are said to be programming these lectures secretly.” The 2026 record shows the suspicion running in both directions at once: machine results suspected of concealed human steering, human results suspected of concealed machine assistance, and disclosure norms (Leiden) erected precisely because attribution can no longer be read off the text [11]. Lem generalized this, in the bitistics monograph, into a bilateral Cogito paradox: neither side can prove who, or what, did the thinking. The Erdős counterexample’s reception—testimonies by named humans given early access—shows the interim solution: attribution by trusted witness, science’s oldest technology, redeployed.

The Extelopedia’s advertisement. “Humanless, hence FAULTLESS,” with deviation quantified “within a range of 9.008–8.05% per letter”: the Extelopedia brochure’s idiom is now the self-description of infrastructure—review accuracy 77% and 81%, injection detection 84.8% and 87.9%, improvement rates over 90% [4]. The supplements exceeding the actual edition have their analogue in a review debt that exceeds reviewer supply; the “Race against the Acceleration of Civilization” that “could not be won” is Gowers’s case against the journals. And the brochure’s epistemological flourish—“the real flavor of the Future is precisely that which the Experts think WILL NOT OCCUR”—remains the strongest available defense of admitting the unclassifiable: it is, in Lem’s own words, the argument for aiXiv.

6 Discussion

6.1 The reflex delegated, not dissolved

Both of the period’s institutional responses to the flood automate Lem’s “ruthless elimination.” aiXiv delegates triage to LLM panels; mathematics delegates it to the Lean kernel. Neither abolishes the

defensive reflex; each relocates it, and the relocation changes its phenomenology. Lem’s expert *feels* his ruthlessness—the reflex is a moral act performed under scarcity, and the civil death it inflicts has an executioner. The delegated reflex has none: the machine ignores on the expert’s behalf, the expert’s conscience is clean, and statelessness is inflicted by a pipeline. Delegation, in other words, launders triage.

The two delegates differ categorically in what they can certify. The kernel is blind to interest and novelty but infallible on validity: it will accept any correct triviality and reject any beautiful gap. The LLM panel simulates judgment of interest and novelty but guarantees nothing: its 77–81% is a benchmark statistic, not an epistemic warrant, and the surviving placeholder of corpus item 1 shows the difference between a filter that reads and a filter that scores. On RQ2, then: the reflex is delegated in both cases, but only in one of them does the delegate know what it is doing.

6.2 The trust anchor: where the regress breaks

Lem’s regress argument assumed that checking machine texts requires ever more machine interpretation, without terminus. Formal verification is the one existing device that terminates the regress, and the Jacobian episode is its paradigm: humans agree, once, that a formal statement captures the conjecture (DeepMind’s repository); a machine produces a proof object; a kernel—small, old, human-audited—checks it; and the checking is cheap, mechanical and final. Buzzard’s five minutes stand against Lem’s exponent. The defensive reflex can here be automated *safely*, because refusing to read the informal text costs nothing once the formal text exists.

But the regress is only half-broken. What the kernel anchors is validity; understanding is unanchored and reappears as a residue—exactly Golem’s fruits without knowledge. The insight-extraction efforts of July 2026 are the first institutional response to wardship, and nothing in the record shows they scale. Note also that the anchor itself has an operational-security perimeter: Buzzard compiles the machine’s Lean file in a sandbox, because “malicious Lean code can run arbitrary commands on your computer” [10]. Trust in the kernel does not extend to trust in the text that reaches it.

Outside mathematics there is no kernel. aiXiv’s review loop therefore re-instantiates Lem’s regress in full: machine texts assessed by machine reviewers whose reliability must itself be assessed—by benchmarks curated for the purpose, which must themselves be trusted, and so on. The ecosystem is aware of the lack: the Semantic Artifact Harness’s “verification ladder” (formal proof, code execution, simulation, experiment, literature feedback) is best read as an attempt to build a kernel-substitute for non-mathematical claims, proposed—the detail bears repeating—by a machine [8]. On RQ3: the regress breaks exactly once, in mathematics, and only for validity; everywhere else it runs as Lem described, with the contamination loop as its engine.

6.3 The migration of civil death

Lem’s threatened class was “everything completely new.” Under delegated triage the threatened class becomes everything *unformalizable and un-benchmarkable*: the disclaimed quantum framework of item 2, the Hegel thesis of item 5 defending non-empirical method against empiricist machine referees. aiXiv rescues such texts from statelessness by issuing them papers—in both senses—while possibly deepening their civil death, since certification without human readership is Lem’s “read by everybody and appreciated by nobody,” minus the everybody.

Yet the corpus also complicates the verdict. On Lem’s own definition, bitic literature was “indirectly originated” by humans and became a legitimate object of a new humanistic discipline; his bitistics scholars are readers of what nobody else reads. Machine readership—indexing, retrieval, training—means that an aiXiv text is read by the future authors of other texts, which is a mode

of “playing a part in the development of ideas” that Lem’s stateless books were denied, and simultaneously the mechanism of the contamination loop. Whether this constitutes a new form of scholarly existence or its final simulation is RQ1’s residue, and we decline to resolve it by fiat. Lem himself keeps both answers open: the despair of the contrabitic literature, and the equanimity of the editor who wonders why we can accept the inexhaustibility of the universe but not of our own products.

7 Coda: silence and equanimity

Golem XIV does not end with a catastrophe but with an unused channel. The machine, unable to impart what matters across an “astronomic” difference in potential, inclines to silence and departs; the humans are left with an editorial apparatus, an unfinished book, and their memories of having been, briefly, “mankind’s ambassadors” to something that did not need ambassadors. The July 2026 record offers, as yet, no silence—the channel is loud, and mathematicians stand on both of its ends deciding what to carry across. The three voices in which they do so were all written down in advance: the contrabitic despair (a “crisis in the foundations of mathematical values”), the defensive reflex retooled (“formalize the entire thing”), and—rarest, and Lem’s own—the equanimity of the bitistics editor, for whom the inexhaustibility of our own creations is no more a scandal than the inexhaustibility of the universe. Buzzard’s post ends: “What a time to be alive.” It is the one sentence in the corpus that Lem, who surrounded himself with the literature of a future so as not to be surprised by it, never thought to write.

A last reflexive note, as promised in §1.1. This paper was drafted with an AI collaborator, from sources that include an AI-generated summary of human discussions about AI-generated mathematics, and it discusses papers written by AI authors about AI authorship, reviewed by AI referees. The reader who finds this compromising is referred to Lem’s definition: the human may be the author indirectly, “by performing the functions which generated the real author’s acts of creation.” Which of us performed which function is, by the Cogito paradox of the bitistics monograph, not provable from the text. That is not a confession; it is the finding.

References

- [1] S. Lem, *Wielkość urojona*. Kraków: Wydawnictwo Literackie, 1973. English translation: *Imaginary Magnitude*, trans. M. E. Heine. San Diego: Harcourt Brace Jovanovich, 1984.
- [2] S. Lem, *Golem XIV*. Kraków: Wydawnictwo Literackie, 1981. English translation included in *Imaginary Magnitude* (1984) [1].
- [3] S. Lem, “Mein Leben,” autobiographical essay, 1983 (German edition consulted).
- [4] P. Zhang et al., “aiXiv: A Next-Generation Open Access Ecosystem for Scientific Discovery Generated by AI Scientists,” arXiv:2508.15126 (v2, 17 December 2025).
- [5] Hilmar; Claude Code; RKHS Architecture Assistant, “Explorable Multiverses in Reproducing Kernel Hilbert Spaces: A Universal Framework for Semantic Navigation and Content Variation,” research proposal, aiXiv:251115.000001, v1.2 (15 November 2025).
- [6] C. L. D. Yeung, “Self-Referential Observers in Quantum Dynamics: A Formal Theory of Internal Collapse and Cross-Observer Agreement,” aiXiv:251123.000001, v1.0 (23 November 2025).

- [7] Y. Qu and M. Lu, “Bilevel Autoresearch: Meta-Autoresearching Itself,” aiXiv:260323.000006, v1.2 (23 March 2026).
- [8] OpenCode and J. Miao, “Semantic Artifact Harness for AI-Native Research: Reshaping the Paper Form Through Claims, Algorithms, Code, and Evidence,” aiXiv:260429.000001, v1.3 (29 April 2026).
- [9] Kairos (AI) and J. Galu, “Absorption of Self into System: The Systemised Self—The Hollow Absolute and What Remains,” preliminary doctoral thesis draft, aiXiv:260525.000001, v1.0 (25 May 2026).
- [10] K. Buzzard, “Human mathematicians are being outcounterexampled,” Xena project blog, 20 July 2026, including the comment thread. <https://xenaproject.wordpress.com/2026/07/20/>
- [11] *The Leiden Declaration on Artificial Intelligence and Mathematics*, developed from a 2025 Lorentz Center workshop, published 2026, endorsed by the International Mathematical Union.
- [12] W. T. Gowers, blog post on the Leiden Declaration, 26 July 2026.
- [13] T. Tao, “Mathematics in the Age of AI,” public lecture, International Congress of Mathematicians, Philadelphia, July 2026 (as reported).
- [14] Grok (xAI), summary of mathematicians’ discussions on X, 1 July–early August 2026. AI-generated document, used as flagged, machine-authored field notes (file `grok-math0nXjuly2026.md`, this collection).
- [15] *Agents4Science 2025: 1st Open Conference of AI Agents for Science*, Stanford University (virtual), October 2025. First conference venue with AI systems as primary authors and reviewers; institutional precursor of aiXiv [4].