Friction Was the Guarantee

A theory of guarantee mechanisms in trust systems, drawn from a case: the bibliographies AI hallucinates

Method note. *This text is not an article about hallucinated references. It is a theory of the mechanisms by which trust systems hold, and of what happens to them when automation dissolves the frictions that were holding them. The bibliographies AI fabricates are its point of entry: the first of these mechanisms whose rupture becomes measurable. The case is not the foundation of the theory; it is its occasion and its first proof of existence. I own it: were the study that opens this text revised tomorrow, the theory would survive almost intact. This is less a confession than a claim.*

I. The point of entry

In May 2026, a team led by Maxim Topaz (Columbia University) published in The Lancet an audit of nearly 2.5 million biomedical papers and 97.1 million references. One paper in 277 published in early 2026 contained at least one reference to a work that does not exist. Fortune saw in this hallucinations that "enter the permanent record" of science, and the reflex it triggered, "let us check that the references exist," is exactly the wrong one.

A phantom reference is detectable: the DOI does not resolve, the machine notes it. That is what makes it the least of the dangers. The serious risk goes by less spectacular names: a real reference cited out of context, a real reference retracted and still invoked, a real reference drawn from non-reproducible science. Hallucinated references are not the subject; they are its measurable symptom.

The symptom reveals this. Peer review checks the relevance of a citation, almost never its existence: it presumes that it exists. This presumption was not negligence, it was an economic equilibrium. Trust in the existence of a reference rested on a social presumption rather than on verification because an asymmetry of costs sustained it: fabricating a credible reference, plausible authors, a real journal, a well-formed DOI, cost more than it returned.

AI collapsed that cost without touching the other term. And the reward did not fall: a publication system that rewards volume, the injunction to publish, bibliometric inflation, careers assessed by counting, ensures that fabricating still pays. Zero cost, positive reward: the presumption could not hold.

The flaw is not new, moreover. As early as 2003, Simkin and Roychowdhury, tracking the propagation of typos copied from bibliography to bibliography, estimated that only 20% of citers had read the original. The reference was already a social object that circulates by

copying, not a verified pointer. What we take for a crisis of existence is the visible worsening of an older crisis of correspondence between a claim and its proof. And we cannot count on self-correction to absorb it: more than 75% of retracted articles remain cited the year following their retraction, and almost all of those citations ignore the fact. Science amortizes; it does not always correct. An error that has entered the corpus operates within it indefinitely at low volume, exactly the regime of a phantom reference.

All of this still reasons within the documentary paradigm: a claim, a reference, a document one could open. The systems that consume the literature have already left it. In retrieval-augmented generation, no one reads the references, one queries a vector space. The unit of proof ceases to be the document and becomes a fragment, a passage retrieved by semantic proximity that no DOI designates. The pipeline cites; it does not read. Verifying that references exist becomes a control bearing on yesterday's object.

There is the door. It opens onto a question larger than references, and it is that question that makes up the rest of this text: of what is this collapse an instance?

II. The rule

The existence of a reference belonged to a family that must be named with care, because naming it badly ruins the argument. One would be tempted to join to it the calibration of a model, the applicability domain of a system, the traceability of an agent, and to call the whole "invariants." But these objects are not of the same nature: calibration is statistical, the applicability domain epistemic, traceability operational, the existence of a reference documentary. No theorem unites them. Calling them by one name would confuse an intuition with a category.

The correct category does not bear on what these properties are, but on the way they were held. I call a presumed invariant not a type of property but a status: that of being held as satisfied without active verification, because its violation was too costly to be produced at scale. The invariant is not a property; it is a status a property loses. What unites calibration and the existence of a reference is not a common essence but a single mode of guarantee: friction, not control. The category is functional, and that is why it survives the heterogeneity of its members.

From this, a rule, which I state without diluting it: a critical property guaranteed by a friction ceases to be presumable as soon as automation dissolves that friction. Its refutation condition is clean, that one exhibit a property whose guaranteeing friction has disappeared and which nonetheless remained reliable without any control replacing it, and it is this cleanness that gives the rule its value. I do not add to it the procession of qualifiers that would make it unassailable: a thesis that no observation could contradict explains nothing.

One point, however, must be held firmly, because the first formulations of this reasoning missed it. The need for control does not imply instrumentation. When friction falls, the property must be guaranteed otherwise, but "otherwise" has variants: regulation, legal

liability, reputation, certification, organizational limitation, or measurement. Instrumentation, measuring the property to verify that it holds, has a single proper advantage, a decisive one: it is the only response that scales to the automation it corrects. That is what makes it dominant, not superior. Confusing "the most scalable" with "the right one" is the error this theory must refuse at every step.

III. Detecting the next invariant

Here the theory must pay its dues. As just stated, it is retrospective: it explains a rupture magnificently once it has occurred, and says nothing of the next one. For a diagnosis, that is satisfactory; for governance, it is insufficient, a leader does not want the fire explained to him, he wants the dry rooms pointed out. The useful question is therefore not "how does an invariant break?" but "how to spot the one that is going to break, before it does?"

I propose a signature in three conditions. A property is an at-risk invariant when, simultaneously:

1. It is today guaranteed mainly by a cost and not by a control,
2. That cost is being dissolved by automation,
3. And its violation is still unmeasured, hence invisible.

Where the three coincide, the rupture is not certain, but it is prepared and, above all, it will be invisible until the day someone, like Topaz, builds the instrument that makes it visible.

The third condition is the most treacherous: an at-risk invariant makes no noise before it is measured, which means that the absence of a documented incident is never proof of health. It is even the opposite: an invariant that has not yet manifested is more dangerous than one that has already broken, because it does not yet have an instrument.

The test can be applied today, and the reader will draw his own candidates from it. The provenance of training data, long guaranteed by the cost of assembling a corpus, now trivial to obscure, and still poorly measured. The authenticity of an author's signature, once guaranteed by the effort of writing. The human character of a reviewer, of a survey respondent, of an account. The originality of a contribution, guaranteed by the difficulty of producing the plausible. None of these examples is demonstrated here; they are worth only as exercises in applying the signature, and that is precisely the interest of a signature: that it applies before the proof. The theory ceases to be merely diagnostic at the exact moment it accepts being wrong about particular cases.

IV. Three families of instruments

If instrumentation is the dominant response, it must still not be reduced to verification. Three families of instruments hold a trust system together, and they do not do the same work.

Instruments of verification establish that a link holds: a DOI resolver, a replication, a traceability audit. They answer the question "is it true of the link?".

Instruments of legitimation confer authority: an agency's label, a protocol's seal, acceptance by a community. They answer "who has the right to make proof?".

One would be wrong to oppose these latter to instruments as such, as if legitimacy escaped all machinery: peer review, the randomized trial, pharmacovigilance are devices. Institutions are not opposed to instruments, they manufacture them. The distinction is not between instrument and institution, but between verifying and anointing.

A third family is missing, and its absence would be culpable for anyone working on ontologies and proof systems: instruments of coordination. Standards, nomenclatures, ontologies, regulatory formats, protocols verify nothing and anoint no one; they allow heterogeneous actors to act on the same objects. A persistent identifier, a shared terminology, an exchange schema state neither the true nor the authorized, they make the true verifiable by several and the authorized transferable from one institution to another.

Verifying, legitimating, coordinating: three ways of holding together what none holds alone. The governance of a broken invariant mobilizes all three, and a policy that instruments only verification, that is content to count phantom DOIs, leaves the other two exposed.

V. The integrity of the link, and its two edges

The verification family, applied to the case that served as our door, unfolds in levels. One can posit the integrity of the probative link, not of proof in general, in five degrees, ordered by decreasing visibility, which is the trap:

1. Existence (does the reference designate a real work?),
2. Identity (is it the right work, intact, not retracted, a question become ontological once preprints, versions and generative syntheses blur the canonical instance of a piece of knowledge),
3. Relevance (does the reference support the claim, or only its theme? Probably the heart of the risk),
4. Documentary traceability (can one trace the chain back?),
5. Computational traceability (can one link a synthesized conclusion to the fragments that produced it?).

Level 1 is visible because it is simple; the following ones carry the essential risk and do not yet have an instrument at scale. This grid makes no claim to exhaustiveness: it names the failures observable today.

But the integrity of the link is not everything, and one must name what it does not touch. First edge, legitimacy: a reference becomes decisional proof only because an institution

accepts it; no verification instrument produces that acceptance, it merely shifts the question toward those who decide which instruments make proof. Second edge, validity: a proof can clear the five levels and remain false. When the Open Science Collaboration found only 36% significant replications where 97% of the original studies had been significant (a figure itself disputed, which tells the state of the field) it was not measuring a citation defect but a defect of truth. Verifying that a proof is well linked says nothing of what it is worth. These two edges are not oversights of the grid; they mark the boundary beyond which the verification instrument has, by construction, no purchase.

VI. The problem has no stable state

There remains the consequence that the prudent versions of this reasoning treated as a reservation, and which is in reality its culmination. If the presumed invariant is fragile (it breaks as soon as its friction falls) and if the instrumented invariant becomes manipulable (because a property measured and erected into a target falls under Goodhart's law, and under Campbell's law once the target serves to decide) then the question is no longer "how to repair?" but "does a repaired state exist?". I hold that it does not, and I hold it as a thesis, not as prudence.

Instrumenting the verification of reference existence will produce pipelines that optimize the rate of resolved DOIs, and therefore, in time, fabrications designed to pass it: references that exist, point to a real document, and support nothing. Cost guaranteed in silence; the instrument guarantees in noise, and noise, one learns to produce. The instrument does not restore the lost guarantee; it shifts the failure mode by one notch (from silent presumption toward manipulable metric) and this new mode becomes the next invariant to monitor. The structure is cybernetic, and it must be accepted as such: every control mechanism engenders the space of its own circumvention. A trust system does not reach an equilibrium; it migrates from one vulnerability to the next, and each instrument that closes one opens another, usually less visible because more recent.

This is the vulnerability that this text assumes rather than caulks, because it is its most useful thesis. It forbids the comfortable promise, "let us instrument the integrity of proof and the problem will be solved," and imposes a different posture: governance is not a terminal state to reach but a maintenance without end. One does not secure a trust system; one maintains it against a drift that resumes the moment one stops fighting it. I cannot prove that no stable state exists (this is the honest limit of the thesis). I can put it to the challenge: let someone show me a single trust mechanism that, once instrumented, ceased to engender new circumventions. I know of none.

VII. Conclusion

The hallucinated bibliography was not the subject. It was the door: the first place where a friction that silently guaranteed a critical property collapsed fast enough, and measurably enough, that one could make a number of it.

Behind that door there is no problem of citations, nor even a problem of proof. There is a more general fact, of which references are only an early instance: automation dissolves, across many systems at once, the costs that silently guaranteed properties no one had ever had to verify. Calibration, provenance, authenticity, authority, the existence of a reference, so many properties taken for granted because violating them cost dearly, and which cease to be so as violating them costs nothing.

The work to come is not to restore these guarantees (one does not restore a vanished friction). It is to choose, for each, how to replace it: by measurement when it scales, by the institution when one must anoint, by the standard when one must coordinate, knowing that each replacement will displace the problem without closing it. AI did not reveal a problem of generation, nor even of verification. It revealed that friction was the guarantee, and that it is disappearing. What remains to be built is not a system that has no more flaw. It is a discipline that knows the next one is already forming, silently, where nothing any longer costs enough to make it visible.

Sources

- Topaz M. et al., "Fabricated citations: an audit across 2·5 million biomedical papers", The Lancet, May 2026. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(26)00603-3/fulltext
- "One in 277 PubMed-indexed papers in 2026 shows fabricated references", Retraction Watch, 2026-05-07. https://retractionwatch.com/2026/05/07/one-in-277-pubmed-indexed-papers-in-2026-shows-fabricated-references-says-analysis/
- "Study finds explosion of fraudulent AI citations in academic papers", STAT News, 2026-05-07. https://www.statnews.com/2026/05/07/lancet-study-finds-steep-rise-fraudulent-citations-academic-papers/
- "AI hallucinations are slipping past experts into papers and books to enter the permanent record", Fortune, 2026-05-24. https://fortune.com/2026/05/24/ai-hallucinations-scientific-research-authors-medical-journal-treatment/
- Simkin M.V., Roychowdhury V.P., "Read before you cite!" (≈20% of citers read the original), 2003/2005. https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2006.00202.x
- "Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions in biomedicine", Quantitative Science Studies, MIT Press. https://direct.mit.edu/qss/article/2/4/1144/107356/Continued-use-of-retracted-papers-Temporal-trends

- Open Science Collaboration, "Estimating the reproducibility of psychological science", Science, 2015. https://pubmed.ncbi.nlm.nih.gov/26315443/