

# Science that Compounds: The Need for A New Substrate for Research in the Age of AI

François Lanusse<sup>1</sup>✉ and Liam Parker<sup>2</sup>✉

<sup>1</sup> Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM

<sup>2</sup> University of California, Berkeley

## CORRESPONDENCE

✉ francois.lanusse@cnr.fr

📍 François Lanusse

✉ lharker@berkeley.edu

📍 Liam Parker

## KEYWORDS

agentic AI; open science;  
research transparency;  
reproducibility; provenance;  
scientific software

## INITIATIVE

This paper is a perspective  
from [LightCone Research](#), an  
open-source initiative building  
tooling for scientific research in  
the age of agentic AI.

*“Nullius in verba.”*

Take nobody’s word for it.

— The Royal Society

Modern science rests on the basic principle that what enters the scientific canon must be trustworthy enough to serve as a foundation for further work. While this principle has never been perfectly upheld, the enterprise functions because the volume of work has roughly matched the community’s capacity to scrutinize it.

With the emergence of AI, many of the activities that have historically set the pace of research — e.g., combing through the literature, implementing software, tracking experiments, writing up the work — are being compressed. The interval between an idea and a presentable result is shrinking, and it is particularly pronounced in disciplines that are not bottlenecked by data acquisition. Indeed, armed with AI, a single research group can already move faster on its ideas. Multiplied across a field, this amounts to a step-change in the rate at which results can enter circulation<sup>1</sup>.

This compression can lift the constraints on what a scientist’s curiosity can pursue and enhance their ability to tackle ambitious and far-reaching research questions. However, it also risks outpacing the community’s already strained capacity to scrutinize new work, thereby eroding the rigor that makes the scientific canon worth building on<sup>2</sup>. Nonetheless, this same AI revolution is providing us with a

<sup>1</sup>This remains true even without invoking AI-powered fully autonomous research, which is not the point of this paper.

<sup>2</sup>This is not the only risk, and many deep, legitimate questions about the future of science and AI — including how scientists are trained, how work is funded, and how credit is assigned — will take time for the community to settle.

powerful new set of tools, opening a rare opportunity to reimagine how good and rigorous science is produced and shared.

In this paper, we narrow our focus to one fundamental question: what foundations can we lay today such that, as new work accelerates, the scientific canon remains trustworthy enough to build on?

## The bottleneck caused by increased productivity

### VOLUME IS AN ASSET; PER-RESULT TRUST IS A BINDING CONSTRAINT

An immediate concern of increased scientific output is volume — a literature growing faster than anyone can read it. However, volume is not, on its own, what erodes science. Indeed, a searchable literature of a million sound, well-indexed results is an asset. Moreover, the same AI systems that are driving the productivity surge behind this increased growth in scientific literature are also sharpening our ability to search it. Nor does the growth of such a literature damage the incentive structure of science: if anything, it pushes recognition toward results that constitute genuine leaps and away from purely incremental, mechanical contributions, whose marginal value collapses once they can be produced cheaply at scale.

Instead, what erodes science is an increasing volume that cannot be trusted to support further work — and this sort of *per-result* trust is precisely the constraint that today's agentic systems run into hardest. The technology that produces results faster also currently produces results whose individual reliability is difficult to establish, even for the scientist steering the AI, let alone for the community downstream. *The fundamental bottleneck on AI-powered science, then, is the widening gap between what is produced and what can be vetted.*

The mechanism we have today to establish whether a result can be trusted is peer review, which does not ask for correctness in a formal sense (most sciences do not trade in mathematical truth) but for *soundness*: that the analysis is internally coherent, that its conclusions follow from the evidence, and that none of its steps are nonsensical on inspection. A scientific result rests on a chain of choices — what hypotheses to test, which models to fit, which priors to adopt, which data to include, which approximations are acceptable — each defensible but also open to challenge, and none of which reduce to the kind of mechanical verification that a proof assistant like [Lean](#) can perform on a theorem.

Establishing soundness has never been perfect or easy, and the more demanding scientific communities have already been pushing the bar upward for years: multiple reviewers, area chairs and meta-reviews, mandatory source-code disclosure, reproducibility checklists, extensive ablation requirements. These mechanisms have raised the floor but remain imperfect and increasingly time-consuming; they also expose the limits of the format that soundness is currently checked against. In much of science, the effective unit of submission remains a text-first article<sup>3</sup> — with code and data attached as supplements at the author's discretion — which forces the reviewer to reconstruct the analysis from a textual summary. This only works as long as the summary is honest and the details glossed over are mostly routine. Neither of those conditions survives contact with AI. Fabrication, traditionally a real but secondary concern of peer review, becomes the primary concern when the producer of a result is an AI that can confidently generate plausible-looking output that is untethered from any actual computation. Moreover, the details missing from the textual summary are no longer routine boilerplate but the very places where an agent's reasoning may have failed silently.

<sup>3</sup>This is not universal: efforts such as AGU's [Notebooks Now!](#) have already been pushing toward richer units of scientific communication, but these efforts remain unevenly adopted.

Peer review is the most visible place where the form of the record matters, but it is far from the only one. The same need to vet results arises at every step of a scientific result's lifecycle: when a scientist scrutinizes the output of their own AI assistants, when collaborators check each other's work in progress, when a reviewer reasons about an analysis they did not run themselves, and when a future reader returns to a result years later and tries to re-establish that its conclusions still hold. At each of these steps, the format is what determines whether trust can be re-established efficiently or has to be taken on faith. The challenge, then, is to define a form of scientific record whose soundness can be efficiently re-established at every step of its lifecycle.

## The investment that compounds

### A NEW COMMONS FOR THE SCIENTIFIC RECORD

Our answer to the challenge mentioned above is structural: an analysis can be vetted efficiently, by either a human or a machine, when it satisfies three properties.

#### Results are provenance-certified

A record ties every plot, number, and claim back to the data, code, and decisions that produced it. Provenance directly eliminates the most acute failure mode of agentic work, namely fabricated or hallucinated results, and does so without requiring reproduction.

#### Analyses are fully observable

Code and research artifacts, but also every consequential decision — which estimator, which prior, which cutoff, which dataset, which preprocessing step — as well as the reasoning that motivated them are inspectable.

#### Analyses are scientifically legible

The analysis is reorganized around its consequential claims, decisions, and insights, so readers can understand and scrutinize the work at the level that matters while retaining direct paths into the evidence, reasoning, and code behind any point that requires closer inspection.

None of these notions are fundamentally new. Bioinformatics has built much of the substrate for reproducibility and reuse over the past decade — workflow systems like [Snakemake](#), [Nextflow](#), and [Galaxy](#), and shared infrastructure like [WorkflowHub](#). Other communities have followed a similar impulse — experimental particle physics, for example, through CERN's [REANA](#) platform. Standards like [W3C PROV](#) and packaging conventions like [RO-Crate](#) have been pushing in the direction of provenance tracking for years. Observability has been chipped at from different angles, by mandatory source-code disclosure, computational notebooks, and experiment-tracking platforms like [Weights & Biases](#). And the manuscript itself has always aimed to be the scientifically legible account, but is rarely complete. Each of the principles we outline can be implemented today as a matter of good practice. But there is a good reason why they are not ubiquitous: they are simply too time consuming for a typical research team to follow.

*Agentic AI allows us to flip this calculus by directly addressing the very problem it creates.* When the work of producing results is assisted by AI, the marginal cost of also producing the provenance trace, the decision log, and the scientific-level summary collapses toward zero. Documentation, formalization, and observability become things that can be built into the process by construction rather than negotiated against an author's time. The very systems that drive the productivity surge can carry the cost of keeping the canon trustworthy.

Together, these principles fundamentally change how a result can be vetted. Provenance certification eliminates fabricated results by construction – the most acute failure mode of agentic work – without ever requiring the verifier to re-execute anything. Observability and scientific legibility allow a reviewer to navigate from the scientific-level account directly to the choices and code that warrant scrutiny, stopping at the level of detail the question demands.

## A direction worth committing to

### AN OPEN SPECIFICATION FOR THE SCIENTIFIC RECORD

In practice, these principles can be satisfied by a new layer that sits between the code and the paper, whose role is to capture and enforce provenance, observability, and scientific legibility by construction. Concretely, this layer takes the form of an open specification: a contract for what a scientific record must contain, with the right level of standardization to enable a rich ecosystem of tools to grow on top of it. The choice to invest in a specification rather than a library or a platform is deliberate – a specification fixes a contract while leaving implementations open, which is what allows independent groups to build interoperable tools on a shared foundation rather than around any one group's stack. By decoupling this contract from the agent layer, we are left with a form that remains stable as AI evolves – letting the scientific record persist while serving as a fixed anchor around which the tooling can adapt over time. Get the specification layer right, and the layers above it become tractable in a way they are not today. Foremost among them is review at scale – review carried out by humans, AI assistants, or some mixture of the two, focused on the choices that warrant scrutiny while keeping the ability to scrutinize the work in depth without worrying about fabrication.

We are breaking ground on this foundation today by launching the development of the *Agentic Schema for Transparent Research and Analysis (ASTRA)* – an open-source schema for representing scientific analyses. However, the full specification cannot be defined by any single group; it has to be built openly and collaboratively, alongside the working scientists who will be its first users, reviewers, and critics, and kept responsive to the field over the long arc of AI's continued evolution. Although still early and evolving, we intend for this standard to ultimately form the seed for a broader suite of tooling that makes it straightforward for scientists, working with AI, to author and review analyses that satisfy the three properties identified above by construction. The curious reader can learn more about ASTRA at <https://astra-spec.org>.

We believe a bright future for scientific research in the age of AI is still ours to build, and it will be built by scientific communities that come together to construct the tooling needed to harness this transformational technology. Yes, AI will reshape how science is done in material ways, but our outlook is fundamentally positive. In some fields, the step change in research capability stands to unlock concrete benefits for humanity – in health, materials, climate, and elsewhere – that would otherwise be decades away. Across the sciences, future scientists will always have something to be curious about; AI will only empower them to push that curiosity further. But whether the canon they inherit continues

to compound, each result a sound foundation for the next, as it has for every generation of scientists before them, depends on the foundations we lay now. Getting that right, openly and together, is the work in front of us. Let's take it up together.

## Acknowledgments

The ideas in this piece were shaped and sharpened over the past year through conversations with many colleagues. We are particularly grateful to the Lightcone Research team, Alexandre Boucaud, Cail Daley, Kangning Diao, and Nolan Koblishke, and to our advisors Uroš Seljak, Fernando Perez, and Kyle Cranmer. We also appreciate the helpful discussions and feedback from David Spergel, Jean-Luc Starck, and Sebastian Wagner-Carena. The views expressed here, and any errors in them, are our own.