The Replication Engine

How to build automated replication infrastructure for better, faster science | **Abel Brodeur and Bruno Barbarioli**

The Replication Engine

How to build automated replication infrastructure for better, faster science

Abel Brodeur and Bruno Barbarioli

This essay is part of The Launch Sequence, a collection of concrete, ambitious ideas to accelerate AI for science and security.

Summary

The results in many science papers do not reproduce when independently tested. This crisis in the integrity of scientific findings leads to billions of research dollars wasted each year, delayed scientific progress, and misinformed policy and investment decisions. We propose building a comprehensive AI-powered infrastructure that automatically reproduces scientific findings across all computational research fields at the moment of publication, using advanced AI agents to parse papers, reconstruct computational environments, execute analyses, and flag irreproducible results.

This initiative would start with a lean pilot program, costing just $10 million over three years, to create the AI verification system and start its rollout. This pilot would enable the software implementation, the establishment of necessary standards, and their integration with selected partnering publishers and academic institutions. After the pilot confirms the efficacy of such an automated system, it could scale across the entire scientific community without further significant government investment.

The full vision includes systematically auditing computational research across physics, economics, psychology, computer science, climate science, and beyond, creating a verified knowledge graph of human understanding, allowing all research to have a vetted lineage, accelerating the pace of scientific progress, and establishing US research institutions as leaders in scientific integrity.

Motivation

The billion-dollar global reproducibility crisis

Scientific progress rests on a broken foundation that spans every field of human inquiry. The landmark Reproducibility Project in psychology found that between one third and one half of studies could be successfully replicated. In experimental economics, preregistered replications and robustness checks showed that nearly half of celebrated results vanished under scrutiny. In climate science, critical temperature reconstructions have been challenged when code and data became available years later. Physics, despite its reputation for rigor, sees major retractions when computational errors are discovered in cosmology and particle physics simulations.

The reproducibility crisis wastes billions across all research fields, slows the pace of scientific progress, and misinforms action in critical areas. Irreproducible preclinical research misdirects $28 billion in biomedical R&D alone. Failed economic studies lead to misguided policies that cost taxpayers billions. Flawed climate models delay critical environmental responses. Meanwhile, irreproducible engineering simulations slow infrastructure development and manufacturing innovation.

Every irreproducible study actively misleads other researchers across all fields, creating cascading effects that compound damage throughout the entire scientific enterprise and steer crucial decisions in government and industry in the wrong directions. In psychology, the “ego depletion” theory spawned thousands of studies and influenced public policy for decades before systematic replications revealed it was largely false. Similarly influential work on “priming” effects led to costly interventions that likely never worked. In economics, spreadsheet errors misled the policy debate over the debt-to-GDP ratio in the early 2010s. In computer science, irreproducible machine learning claims have misguided entire research programs and startup investments.

The vision: A self-correcting science across all fields

Imagine a world where every computational scientific claim, whether in particle physics simulations, machine learning algorithms, or social science analyses, comes with algorithmic verification. If every robust published computational finding came with a green verification badge, scientists could build on each other's work with confidence.

Automated verification would allow scientific research to move at the speed of discovery rather than being slowed by false leads and post-publication detective work. Funding agencies could direct resources toward genuinely robust findings instead of chasing false leads, whether in fusion energy research, drug discovery, or artificial intelligence.

Science could re-earn public trust, as visible, machine-audited evidence of scientific reliability across all fields becomes available. Costly mistakes could be avoided in the many policy decisions that rely on products of physical and social science. And US institutions of science would hold up the global gold standard for scientific integrity across all disciplines.

Why now: The AI advantage across empirical scientific fields

Previous attempts at large-scale reproduction and replication failed because they required armies of human specialists in each field to manually check codes or recreate experiments. The Many Labs project, despite heroic efforts, managed to replicate fewer than 30 psychology studies over several years. Field-specific reproduction and replication efforts in economics, computer science, and other disciplines face similar scalability limits. Manual reproduction and replication simply doesn’t scale to the 3 million papers published annually across all fields.

Advanced AI changes everything across empirical disciplines. Much science is carried out computationally, e.g., through computer simulations, statistical analyses, and quantitative data analysis — and therefore is suited to AI verification of its methods/reproducibility. Modern language models can parse scientific

papers, extract methodological details from physics simulations to economic regressions, and generate functional code across Python, R, Matlab, Fortran, C++, and dozens of other languages used throughout science. They can automatically provision software environments for everything from climate models to machine learning frameworks, execute complex analytical pipelines across fields, and identify subtle errors that escape human review whether in statistical analyses or numerical simulations.

More importantly, AI agents can work at the moment of publication across all fields, not years later when the damage is already done. They can process the entire corpus of computational human knowledge, not just field-specific samples. And they can continuously improve their detection capabilities by learning from each new paper they process across every scientific discipline.

However, a classic collective action problem stands in the way of developing AI verification systems. Publishers won't build this individually, because benefits accrue across the entire research ecosystem. Field-specific solutions create fragmentation; physics journals won't invest in economics-specific tools and vice versa. We propose a science-agnostic framework that is still able to provide individualized solutions through fine-tuned models for each specific use case. In this scenario, the government can coordinate across disciplinary boundaries while ensuring the infrastructure remains public and nonproprietary. Once the solution is implemented and deployed at scale, all stakeholders would benefit from its use, creating a virtuous cycle: publishers that verify their research have enhanced credibility within their scientific community, leading other venues to adopt the same standards to remain competitive and attractive to submissions and citations until most journals decide to use the framework.

Solution

We envision a system to conduct an automated reproduction of all quantitative research. Such a framework would work in the following way:

- All publishers have access to a cloud-based system hosting the reproduction infrastructure.

- When an author uploads a paper, it's automatically run through that reproduction infrastructure.
- A first AI agent parses the results from the paper to be evaluated. A second AI agent checks that the code the authors submitted runs without issues and produces the results presented in the paper. A third agent checks for coding errors and data irregularities. A fourth agent checks the sensitivity of the main results to reasonable robustness checks.
- The agents assign Green/Amber/Red badges to the components of the paper's computational analysis.
    - Green badges signal full agreement between regenerated output and the paper.
    - Amber badges indicate minor divergences that merit author attention before publication.
    - Red badges flag blocking errors or irreparable gaps in the evidentiary chain.
- The paper is flagged with the results of the AI analysis, and returned to editors and authors, to act on as needed. If the paper is published, the results of the reproduction tests are also transparent to readers.
- Two public knowledge graphs update continuously as audits accumulate: one traces how unverified claims propagate through citation networks, while the other maps collaboration clusters whose work shows unusual fragility.
- Researchers, journalists, funders and investors can explore these visualizations to decide where deeper human replication or additional resources will yield the greatest return.

Phase 1: The Universal Catalyst Program (Years 1–3: $10 million)

The National Science Foundation (NSF) should launch a lean pilot program to provide automated reproduction services across all computational research fields to a diverse consortium of journals.

The NSF should fund this through its existing Directorate for Computer and Information Science and Engineering and designate the entity to run the program after the initial period, if it chooses not to run it itself. Coordinating across all NSF directorates would ensure coverage of physical sciences, social sciences, engineering, and mathematical sciences. The Office of the US Chief Technology Officer (CTO) should provide the initiative with $1 million in cloud computing credits through its existing partnerships.

Implementation

The pilot with major existing publishers would be implemented via a $3 million annual cooperative agreement for technical development over 3 years. This agreement, which could be spearheaded by any academic and/or research institution with experience with large-scale replication, would produce:

- **Technical Infrastructure:** Lightweight, cloud-based AI agents that parse manuscripts across fields, reconstruct computational environments for any programming language, execute analyses, and generate verification reports
- **Universal Publisher Integration:** Easy integration with manuscript submission systems used across all fields — from Physical Review Letters to Nature to American Journal of Public Health
- **Cross-Disciplinary Standards:** A green/amber/red badge system with criteria adapted for different fields by specialists within their research communities (statistical significance tests vs. numerical convergence vs. algorithmic correctness)

Timeline: This infrastructure would be piloted across a range of scientific fields starting in the first year, via a phased rollout.

- The system would process 2,500 papers in the first year, spanning economics, psychology, computer science, physics, climate science, materials science, and other computationally-intensive fields.
- This would start with 50 journals across different fields in Year 1, expand to 250 journals by Year 2, and then scale via adoption based on demonstrated value.
- The expected outcomes by Year 3 would be:

- 15,000 papers verified across all major computational fields
- Universal verification infrastructure demonstrated and proven
- Clear return on investment evidence
- Global adoption beginning as other nations adopt US-developed open source tools

Federal science agencies: These agencies (NSF, NIH, DOE, etc.) have a catalytic role to play in the early adoption of these systems. They should announce they will accept verification badges for grant reporting starting in Year 2, and require them starting in Year 3 in order to either extend grants or provide new ones to grantees. No new bureaucracy arises; the badge simply occupies a field that already exists in digital-object metadata.

Phase 2: The Universal Verification Vision (Years 4–10: Scaling through network effects)

Once the verification system has proven its value via the three year pilot, adoption spreads naturally. Publishers want reliability advantages. Researchers want verification badges. Universities want to train students in best practices. The system becomes self-sustaining without requiring massive government spending.

After Year 3, the costs of the system can transfer to the distributed network of stakeholders who benefit; publishers pay modest fees for premium features, universities contribute computing resources, international partners share development costs and private companies fund specialized modules for proprietary research.

Recommended actions

- The NSF should issue a $3 million annual cooperative agreement for technical development over 3 years, under the Directorate for Computer and Information Science and Engineering's budget.
- The US CTO's office should provide the initiative with $1 million in cloud computing credits through existing partnerships.

- The OSTP should convene a cross-disciplinary publisher roundtable to establish universal metadata standards.
- All federal science agencies (NSF, NIH, DOE, etc.) should announce they will accept verification badges for grant reporting starting in Year 2.

Further resources

- **Brodeur et al.**, "Mass Reproducibility and Replicability: A New Hope," 2024.

  Comprehensive analysis of reproduction and robustness rates across disciplines.

- **Freedman, Cockburn & Simcoe**, "The Economics of Reproducibility in Preclinical Research," *PLOS Biology*, 2015.

  Quantifies costs of irreproducible research.

- **Chen et al.**, "Evaluating Large Language Models Trained on Code," 2021.

  Demonstrates AI capabilities for automated code generation across languages.

- **Senator Heinrich**, "American Science Acceleration Project RFI," n.d.

  Policy framework for 10x acceleration of scientific progress.

- **Crossref**, "Crossref Technical Documentation," n.d.

  Existing infrastructure for scientific metadata across all fields.

Appendix

Technical implementation details

Universal agent architecture

- **Parser agent:** Fine-tuned language models extract methodological details from papers across physics, economics, computer science, and all computational fields
- **Environment agent:** Reconstructs computational environments for Python, R, Matlab, Fortran, C++, Julia, Stata, and other languages used across science
- **Execution agent:** Runs analytical pipelines across all scientific computing paradigms in sandboxed environments
- **Verification agent:** Compares results using field-appropriate criteria (statistical tests, numerical convergence, algorithmic correctness)
- **Cross-field critic agent:** Trained on analytical errors across all disciplines to identify field-specific and universal problems

Lean infrastructure design

- Cloud-native architecture using existing commercial platforms
- Containerized execution environments for reproducible deployment
- API-first design for easy integration with any journal submission system
- Modular components that can be deployed independently across fields

Cost efficiency measures

- Leverage existing open source scientific computing tools
- Use volunteer graduate student reviewers for quality control

- Partner with cloud providers for donated computing resources during and after the pilot program
- Build on proven containerization and orchestration technologies rather than developing from scratch