Reliability in Scientific AI: From Verification to Reproducibility

Munongedzi Mabhoko, Clarkson University, mabhokm@clarkson.edu

Abstract

Reliability is the currency of scientific credibility, yet modern artificial intelligence research often treats it as an optional aesthetic rather than a structural necessity. In machine learning, we have achieved exponential computational scale but not epistemic stability. This paper argues that reliability in Scientific AI must be reimagined as a systemic property,one that spans from formal verification at the model level to reproducibility and governance at the institutional level. A New Architecture and Roadmap for Scientific AI, I propose a layered conception of reliability that connects mathematical proof, data lineage, workflow transparency, and human stewardship. Reliability is not a proof we obtain once, it is a culture we must design to sustain.

1. Introduction : The Reliability Problem in Scientific AI

Scientific progress depends on reproducibility, yet much of contemporary AI operates like alchemy, opaque models trained on undocumented data, evaluated with inconsistent metrics, and deployed into social contexts that resist formal reasoning. The crisis is not merely technical, it is epistemological. Machine learning systems, especially deep neural networks, are increasingly embedded in high-stakes domains like healthcare and drug discovery, but their internal logic remains unverifiable, and their data pipelines are poorly governed.

Reliability, in this context, means more than low error rates. It demands verifiability (that we can prove a system's behavior matches its specification), reproducibility (that independent actors can replicate results), and accountability (that the provenance of data and models is transparent). Yet these layers have evolved separately. Formal verification

in computer science, documentation ethics in data science, and process governance in the sciences. The question is not whether AI can be made reliable, it is whether we can architect systems that make reliability unavoidable.

2. Proving Reliability from Within

At the model level, reliability begins with proof. An Efficient SMT Solver for Verifying Deep Neural Networks marked a critical turning point in formal verification. By adapting the simplex method for piecewise-linear activations like ReLUs, Reluplex allowed researchers to verify that a neural network satisfies specific constraints, for example, that small perturbations in input images do not change output classifications ("Reluplex" 2017). It provided, for the first time, a way to reason mathematically about systems once thought too complex to audit.

However, Reluplex was limited to relatively small networks. The follow-up, The Marabou Framework for Verification and Analysis of Deep Neural Networks (Katz et al. 2019), extended this idea into a scalable, modular architecture that could interface with diverse solvers and handle industrial models. This contribution was not just technical, it reframed verification as a reusable infrastructure.

These systems illustrate that reliability, at its foundation, is a question of provable correspondence. Can we show, with mathematical certainty, that a model will not violate its safety properties? Yet this notion of reliability, while precise, is narrow. Formal verification ensures that the model behaves as specified, but it says nothing about whether the specification itself was valid, or whether the data it learned from was trustworthy.Verification, in other words, can guarantee internal consistency without guaranteeing external truth.

3. Data Governance : Reliability from the Outside In

If formal verification is reliability from the inside out, then data governance is reliability from the outside in. The works of Gebru et al., Mitchell et al., and Sambasivan et al. diagnose how modern AI systems fail long before training begins.

Datasheets for Datasets (Gebru, 2021) proposed a simple but transformative idea, every dataset should be accompanied by standardized documentation. Its motivation, composition, collection methods, labeling process, and recommended uses. Just as electronics require datasheets for safe engineering, so too should datasets expose their conditions of use. This framework extends to trained models, specifying intended uses, limitations, performance across subgroups, and ethical considerations.

In contrast, we can assess what happens when such documentation and governance are absent. Through field studies across sectors, neglected data work, poor labeling, ambiguous ownership, and hidden feedback loops leads to compounding failures downstream. These patterns are referred to as “data cascades,” where early missteps in collection or curation propagate into model bias, deployment harm, and loss of scientific integrity.Reliability cannot be retrofitted after data has been collected, it must be baked into the epistemic scaffolding of data production itself.

4. Why Verification Alone Fails

Formal verification provides guarantees about function, but not meaning. Data governance provides guarantees about context, but not behavior. Between them lies what might be called the missing middle, the engineering of reliable workflows and institutional practices that connect data, models, and scientific reasoning into a coherent whole.

When we treat verification as a one-time mathematical exercise, we ignore the sociotechnical pipeline that generates errors faster than any theorem can fix. As Sambasivan argues, data cascades are often invisible precisely because teams valorize

modeling and undervalue data work. Verification without documentation is sterile, documentation without enforcement is symbolic.
The challenge, then, is not to choose between formal rigor and procedural accountability, but to unify them under a shared architecture of reliability.

5. Reliability as Architecture

A New Architecture and Roadmap for Scientific AI provides that unifying vision. Science today runs on a pre-industrial operating system, characterized by fragmented data, artisanal workflows, and an N-of-1 model of progress. This solution is architectural rather than algorithmic, to industrialize scientific reliability through four interlocking systems, the Scientific Data Foundry, the Scientific Use Case Factory, Tetra AI, and Tetra Sciborgs.

Each pillar maps directly to the reliability crisis described in the earlier literature:

- **The Scientific Data Foundry** parallels Datasheets for Datasets. It liberates data from proprietary silos, standardizes semantics, and enforces lineage, an operationalized form of dataset documentation.

- **The Scientific Use Case Factory** institutionalizes reproducibility by transforming one-off AI experiments into validated, reusable workflows. This echoes Model Cards, but extends them into living production pipelines.

- **Tetra AI** provides a reasoning layer that connects verified models to their data provenance and workflow semantics, ensuring epistemic consistency, a form of dynamic verification.

- **Tetra Sciborgs** embody the human infrastructure for reliability, cross-disciplinary scientist-engineer teams that implement, monitor, and culturally sustain the system.

Grady's architecture converts reliability from a moral aspiration into a design constraint. It is, in effect, a formal verification of science itself not of neural networks alone, but of the institutions that produce them.

6. The Moral Infrastructure of Reliability

Every locked dataset prolongs human suffering. This is not hyperbole, it reframes inefficiency as ethical failure. If data silos delay the discovery of cures, then reproducibility is not just a technical virtue, it is a humanitarian one.

This argument resonates with the ethos of Datasheets and Model Cards, which demand transparency not for bureaucracy's sake but for justice. When researchers fail to document or verify, they obscure accountability for harm. Scientific AI must replace secrecy with stewardship.

Reliability, then, is both a moral and structural imperative. It requires systems that make good behavior the path of least resistance, where provenance tracking, model validation, and open ontologies are automatic, not optional.

In that sense, the Foundry's lineage protocols and the Factory's embedded validation loops act as ethical circuits, they prevent harm not through oversight, but through design.

7. From Formal to Epistemic Verification

Traditional verification asks, "Did the model compute correctly?" Epistemic verification asks, "Did we reason correctly about the world?" The former is mathematical, the latter is philosophical. Yet in scientific AI, both are required.

Datasheets and Model Cards verify procedural transparency. Semantic reliability ensures that the reasoning layer connecting data, workflows, and models preserves the causal and conceptual relationships inherent in scientific knowledge.

In practice, this means an AI system that not only predicts but understands the structure of its own reasoning, able to explain not just outcomes but why those outcomes make sense within the experimental ontology. This aligns with the broader shift toward interpretable and grounded AI, where reliability is measured not just by consistency but by comprehension.

8. Toward Continuous Reproducibility

In high-stakes domains, reproducibility cannot remain a one-time event, it must become continuous. This idea parallels DevOps in software engineering and MLOps in machine learning, but it goes further by embedding compliance and validation as intrinsic steps. Reproducibility becomes a living process, not a post-hoc audit.

Continuous reproducibility is the antidote to data cascades. It transforms documentation from passive record-keeping into active feedback. As more data flows through the Foundry, the Factory's models improve, generating a compounding cycle of reliability.

Here, reliability ceases to be about replication alone. It becomes about inheritance, each iteration inheriting and improving upon the reliability of the last.

9. Reliability as Architecture

Synthesizing across these sources, we can conceptualize reliability in Scientific AI as a five-layer architecture. Reliability, therefore, is not reducible to any single layer. It is the alignment of all five. Formal verification ensures correctness, data governance ensures integrity, workflow reproducibility ensures persistence, organizational stewardship ensures continuity, and moral awareness ensures purpose.

When these layers interlock, AI ceases to be an opaque artifact and becomes a transparent participant in the scientific process.

10. Conclusion

True reliability in Scientific AI will not emerge from incremental policy tweaks or a patchwork of ethical add-ons. It requires a fundamental reconstruction of how we build, document, and reason with intelligent systems. The intersection of verification, data governance, and industrial-scale reproducibility is not merely a procedural convergence, it is the blueprint for a new scientific order. Verification ensures that systems behave as promised, governance ensures that their knowledge is traceable and just, industrialization ensures that reliability becomes self-reinforcing rather than episodic. When these dimensions are designed to interact and not coexist, reliability transforms from a static goal into a living property of the system. The crisis of reproducibility will not be solved by better checklists or stricter reviews. It will be solved when scientific practice itself evolves, from an artisanal craft of isolated models to an engineered ecosystem of compounding intelligence. Verification must cease to be an event that happens after discovery and instead become a culture that defines how discovery happens. If we architect our systems to compound reliability, we architect our science to compound truth.

References

Gebru, Timnit, et al. “Datasheets for Datasets.” *Communications of the ACM*, vol. 64, no. 12, Dec. 2021, pp. 86–92. ACM Digital Library, doi:10.1145/3458723.

Grady, Patrick. “From First Principles A New Architecture and Roadmap for Scientific AI.” *Unvarnished*, 23 Oct. 2025, unvarnished.substack.com/p/from-first-principles-a-new-architecture.

Katz, Guy, et al. “Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks.” *Computer Aided Verification (CAV)*, edited by Rupak Majumdar and Viktor Kuncak, Springer, 2017, pp. 97–117. doi:10.1007/978-3-319-63387-9_5.

Katz, Guy, et al. “The Marabou Framework for Verification and Analysis of Deep Neural Networks.” *Computer Aided Verification (CAV)*, Springer, 2019, pp. 443–452. doi:10.1007/978-3-030-25540-4_26.

Mitchell, Margaret, et al. “Model Cards for Model Reporting.” *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, Association for Computing Machinery, 2019, pp. 220–229. doi:10.1145/3287560.3287596.

Sambasivan, Nithya, et al. “Everyone Wants to Do the Model Work, Not the Data Work: Data Cascades in High-Stakes AI.” *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI ’21)*, Association for Computing Machinery, 2021, pp. 1–15. doi:10.1145/3411764.3445518.