NEOVERITY GROUP

PHYSICAL AI · INDEPENDENT VERIFICATION

You Cannot Grade Your Own Exam

Why Physical AI Needs Independent Verification

Mike Vildibill

Founder and Principal, NeoVerity Group

July 2026

SUGGESTED CITATION

Vildibill, M. (2026). *You Cannot Grade Your Own Exam: Why Physical AI Needs Independent Verification*. NeoVerity Group whitepaper. https://www.neoverity.com/reports/you-cannot-grade-your-own-exam/

Executive Summary

Artificial intelligence is crossing a threshold. For the past several years, its most visible successes have lived in the world of language: drafting, summarizing, coding, reasoning over text. Increasingly, AI is being asked to act on the physical world: to propose molecules, accelerate simulations, steer experiments, tune industrial processes, and inform engineering and safety decisions. This is the domain we call Physical AI, the convergence of machine learning with the long-standing mission of high-performance computing. It is broader than robotics, and it requires no embodiment. What it requires is correctness, because when AI acts on the physical world, being wrong stops producing bad answers and starts producing bad outcomes.

This paper makes one central argument: **a system cannot corroborate itself.** Fluent, confident, internally consistent output is not evidence of correctness, and no amount of scale changes that. The mathematics here is old and settled. Ensemble error decomposes into bias, variance, and covariance; adding more copies of the same estimator drives down the variance term but leaves the covariance term untouched (Ueda & Nakano 1996). Correlated errors cannot be averaged away. A single model, reasoning under one set of weights, shares its own blind spots, so its agreement with itself certifies confidence, not correctness. A July 2026 preprint from Pacific Northwest National Laboratory puts the point sharply, and the modern empirical literature on AI self-correction and self-evaluation confirms it in practice.

The answer is not less AI. It is an independent verification layer that uses distinct, uncorrelated approaches to cross-check a claim before it touches the physical world. The differences in the approaches make their agreement informative rather than self-serving. Independence is the active ingredient. Stacking more models that share training data, architectures, or a common foundation model is the common-cause failure the underlying theorems warn against, not a verification strategy.

None of this is novel doctrine. Everywhere the cost of being wrong is physical, engineering practice already mandates independent verification: IEEE Std 1012 defines it by technical, managerial, and financial independence; NASA operates a permanent IV&V program on exactly those terms; functional-safety standards scale the required independence of the assessor with the severity of the stakes; nuclear regulators re-run the safety case with their own codes; and science itself confers validity through independent replication. Extending this posture to Physical AI is continuity, not invention.

For leaders building or buying Physical AI, the implications are concrete. A bigger model is not a verification strategy. Vendor verification claims should be evaluated on independence, not accuracy alone. The verification layer belongs where the physical stakes are highest, applied before expensive or irreversible action. And someone in the organization must be structurally positioned, and structurally protected, to deliver the independent “no.”

You cannot grade your own exam. In the physical world, the cost of a wrong grade is not a wrong answer. It is a wrong outcome.

1. The Stakes: From Bad Answers to Bad Outcomes

Science has expanded its methods over centuries: from empirical observation, to theory, to computational simulation, and most recently to data-intensive discovery, the progression Jim Gray memorably framed as a fourth paradigm (Hey, Tansley & Tolle 2009). Each expansion brought new power and demanded a new discipline for knowing when its results could be trusted. Theory demanded proof. Experiment demanded controls and replication. Simulation demanded verification, validation, and uncertainty quantification. As AI now enters the scientific and engineering workflow, the same question returns in a new form: what is the discipline by which we know a machine-generated answer is right?

The question is urgent because of where AI is going. Large language models transformed what machines can do with humanity's symbolic knowledge. The next frontier is AI grounded in the physical world itself: systems that combine machine learning with simulation, numerical methods, observational data, and domain constraints to model, predict, optimize, or control real-world phenomena. We call this Physical AI, and we understand it as the convergence of AI with the long-standing mission of high-performance computing. It is emphatically broader than robotics. It includes learned surrogates that accelerate expensive simulations, hybrid AI-plus-simulation design loops, digital twins, inverse problems, and scientific foundation models. No embodiment is required; what defines the category is that the system is computationally engaged with physical reality, not only with descriptions of it.

A simple framework organizes what follows. **Layer 1 is the world**: the physical substrate where consequences are real, from materials and molecules to cells, grids, machines, and engineered and natural processes. **Layer 2 is models of the world**: mechanistic simulation, solvers, learned surrogates, probabilistic estimators, the machinery that predicts physical behavior. **Layer 3 is descriptions of the world**: language and the symbolic layer, what an LLM generates. Historically, HPC addressed Layer 2 in service of Layer 1. LLMs transformed Layer 3. Physical AI is the convergence zone where these layers begin to reinforce one another.

The methods making this convergence concrete are real, cited, and advancing. Physics-informed neural networks embed governing equations directly in the training loss (Raissi, Perdikaris & Karniadakis 2019). Neural operators such as the Fourier Neural Operator and DeepONet learn mappings across entire families of differential equations (Li et al. 2021; Lu et al. 2021). Universal differential equations embed learnable terms inside known mechanistic models, keeping the physics and learning only what is missing (Rackauckas et al. 2020). We are explicit about maturity: these methods are emerging, not turnkey. Reported accelerations, sometimes orders of magnitude over classical solvers, hold within the training distribution and are author-reported. And these are hybrid methods by construction. Physical AI augments simulation; it does not replace it. The field itself says as much: the National Academies' consensus study on digital twins names verification, validation, and uncertainty quantification as a foundational research gap, the open problem on which trust depends (NASEM 2024).

Why does this demand a distinct verification discipline now? Because the cost profile of error changes at Layer 1. In Layer 3 knowledge work, a confident but wrong output costs a rewrite, an embarrassment, a wasted afternoon. In Layer 1 domains, it costs a failed experiment, a flawed design entering certification, an unsafe control action. The 2018 Tempe, Arizona collision investigated by the National Transportation Safety Board is a sobering marker. The investigation found a multi-causal system failure, including operator inattention, disabled emergency braking, and safety-culture deficiencies; among its central factors, the

automated driving system detected the pedestrian seconds before impact but cycled among classifications, object, vehicle, bicycle, and never correctly classified her as a pedestrian in time to act (NTSB 2019, HAR-19/03). The lesson is not that AI is uniquely dangerous. The lesson is that when machine perception and decision are coupled to physical action, classification error becomes physical consequence, and the surrounding system must be engineered accordingly.

The pattern is not confined to vehicles, and not to robotics. In 2021 researchers published the first large independent external validation of a proprietary sepsis-prediction model that was already running in hundreds of U.S. hospitals, a model whose marketed performance had been evaluated by no party other than its vendor. Measured against more than 38,000 hospitalizations, its ability to discriminate sepsis, an area under the curve of 0.63, fell well below the 0.76 to 0.83 the vendor had reported, and it missed roughly two thirds of sepsis cases while generating alerts on 18 percent of all hospitalizations (Wong et al. 2021). This is clinical predictive analytics rather than physics-grounded modeling, but the lesson transfers precisely. The gap between self-reported and independently measured performance stayed invisible until someone outside the vendor was finally allowed to grade the exam, and when a model's outputs steer decisions about real patients, that gap is a safety matter, not an academic one.

The pattern is not even confined to machine learning. The Boeing 737 MAX's Maneuvering Characteristics Augmentation System (MCAS) automatically commanded nose-down stabilizer trim when it inferred an excessive angle of attack. The aircraft carried two angle-of-attack sensors, but MCAS read only one at a time and did not require the two to agree, so a single erroneous vane had no independent reading to contradict it. On Lion Air flight 610 in October 2018 and Ethiopian Airlines flight 302 in March 2019, 346 deaths in all, one bad sensor fed MCAS a false signal, and the system trimmed the nose down again and again. The design's only backstop was the assumption that pilots would recognize and counter it, and that assumption was itself never independently tested: the NTSB found that this erroneous-input failure mode was not simulated in Boeing's hazard-assessment validation, and that the pilots' actual responses "were not consistent with the underlying assumptions about pilot recognition and response that Boeing used" (NTSB 2019, ASR-19-01). Two failures of independence compounded: no independent check on the sensor, and no independent test of the safety case that dismissed the risk. MCAS was a deterministic control law, not a machine-learning system, and that sharpens the argument rather than weakening it. The failure of self-certification is not a quirk of neural networks. It is a property of trusting a single, correlated source of judgment, whatever its technology. Learned perception, a statistical clinical model, a rule-based control law: three technologies, one structure.

That is the bar the rest of this paper must clear. When AI acts on the physical world, "plausible" is not a standard. Correct is the standard. So the binding question stops being "can the system produce a plausible description?" and becomes "is the system correct about the world?"

2. The Seduction: Fluency Is Not Correctness

The most dangerous property of modern AI systems is not that they are often wrong. It is that they are wrong fluently. A generated answer arrives grammatical, confident, internally coherent, and formatted like expertise. Every heuristic humans use to gauge credibility in other humans, fluency, confidence, consistency, is present, and none of them tracks truth in a machine-learning system.

Our shorthand for the necessary precision is **description ≠ prediction ≠ understanding**. A system that generates a plausible description of a physical system has not thereby modeled the physical system. A model that predicts a system's behavior within its training envelope has not thereby understood the mechanism, and may fail without warning outside it. These are three different achievements, earned by three different kinds of evidence.

This is not an anti-LLM claim, and we want to be exact about that. Language models are indispensable at Layer 3 and increasingly useful at Layer 2: specifying experiments, navigating literature, orchestrating simulation pipelines, interpreting results. The point is not that language models are weak. The point is what a plausible description warrants: a hypothesis, not a conclusion.

The evidence that fluency and correctness come apart is strongest exactly where Physical AI lives. Physics-informed neural networks, the flagship method for embedding physics in learning, fail on only moderately harder physics, and the failure is structural: Krishnapriyan et al. (2021) showed that PINNs that learn good models for relatively simple problems can fail to learn the relevant physical phenomena on slightly more complex ones, not because the network lacks capacity but because the training setup itself makes the loss landscape hard to optimize. The failure is invisible in the fit quality. Nor can we simply ask the model when it is unsure: uncertainty quantification for scientific machine learning is an open research problem, with error arising from noisy data, limited data, hyperparameters, overparametrization, optimization, and model misspecification all at once (Psaros et al. 2023). And the confidence a network reports is not a probability of being right: modern deep networks are systematically miscalibrated toward overconfidence, and growing capacity made calibration worse, not better (Guo et al. 2017).

Leading researchers have drawn the same distinction from the architecture side. LeCun (2022) argues that intelligence in the physical world requires learned internal world models that predict and plan, a research program rather than a demonstrated capability, and we cite it as framing. But the framing matters: the distinction between modeling the world and modeling descriptions of the world is live at the highest levels of the field, and it is the correct distinction.

The practical conclusion is uncomfortable and important. In Physical AI, the model's own outward signals, fluency, confidence, self-consistency, carry almost no information about correctness. If those signals cannot be trusted, correctness has to be established some other way. The obvious candidate is corroboration: check the answer. The next section shows why who does the checking is the entire question.

3. The Corroboration Problem: A System Cannot Corroborate Itself

Suppose a model produces an answer and we ask it to check its own work. Suppose it agrees with itself, even across many samples and phrasings. What have we learned?

A July 2026 preprint from Pacific Northwest National Laboratory puts the point sharply: "A single model evaluating every domain under one set of weights produces correlated errors; it cannot mathematically corroborate itself at any scale" (Choudhury et al. 2026, preprint). The formulation is crisp, and we cite the preprint for its articulation. The proof weight, however, does not rest on any single recent paper. It rests on decades of peer-reviewed statistics, and the underlying result is about as settled as results get.

The mathematics is worth stating plainly, because it is the intellectual spine of this paper. The expected squared error of an ensemble of M estimators decomposes exactly into three terms: bias squared, plus variance scaled by 1/M, plus covariance scaled by (1 − 1/M) (Ueda & Nakano 1996). Adding members drives the variance term toward zero. The covariance term does not shrink with M at all. If the members' errors are perfectly correlated, the ensemble's error equals the single model's error, and the aggregation buys nothing. Correlated error cannot be averaged away; it can only be checked from outside the correlation. The same law appears across the field's foundational results. Breiman (2001) proved that the generalization error of a random forest is governed by the strength of the individual trees and the correlation between them, and introduced randomization precisely to decorrelate the members. Krogh & Vedelsby (1995) showed that ensemble error equals the average member error minus the members' diversity; zero diversity, zero gain. And metrology settled the general case long before machine learning existed: repeated measurement averages away random error but leaves systematic error untouched, which is why a device cannot calibrate itself and calibration is defined against an independent reference (JCGM 100:2008; NIST/SEMATECH).

The floor is worth stating in closed form. Average M estimates whose errors share a common variance σ^2 and a mean pairwise correlation ρ. The error variance of the average is then $\sigma^2[1 + (M - 1)\rho]/M$. When the errors are independent ($\rho = 0$), this is the familiar σ^2/M, vanishing as the ensemble grows. When any correlation is shared ($\rho > 0$), it tends not to zero but to $\rho\sigma^2$: an irreducible floor set by the correlation alone. This is not a new theorem; it is elementary algebra on the covariance term above (Ueda & Nakano 1996), and the same variance-covariance structure governs the accuracy of aggregated human judgment (Davis-Stober et al. 2014). But it compresses the argument of this paper into one line. Scale changes M. It cannot change ρ. A bigger stack of correlated checkers cannot cross the floor its correlation sets.

A model checking its own output is the $\rho \to 1$ limit of this expression, the zero-diversity case where the floor sits exactly at the single model's error. Its "second opinion" is generated by the same weights, the same training distribution, the same inductive biases that produced the first, so it adds no independent observation, and an opinion that adds no independent observation cannot move the floor. Self-agreement therefore certifies consistency, which the system was always going to exhibit, and not correctness, which is the thing we actually need.

Precision requires a qualification, and we state it in our own voice rather than the preprint's. "At any scale" should not be read as "self-checking is worthless." A self-critique pass can catch random slips, an arithmetic error, a dropped constraint, and sampling-based self-consistency measurably helps against variance-type mistakes. What self-checking cannot do is remove correlated, systematic error, because those errors are shared between the generator and its self-appointed checker. The achievable self-corroboration is bounded by the residual correlation of a single weight set. Correlated is not identical, but it is not independent either, and independence is what corroboration requires.

The empirical record in modern AI matches the theory. Huang et al. (2024) examined intrinsic self-correction in large language models and found that "LLMs struggle to self-correct their responses without external feedback, and at times, their performance even degrades after self-correction." When improvement does occur in self-correction pipelines, it traces to external signal entering the loop, which is precisely the theory's prediction. And the pattern is not confined to language models. Machine-learning-based science at large has been grading its own exam: Kapoor & Narayanan (2023) documented data leakage producing over-

optimistic, irreproducible results across 294 papers in 17 disciplines, published findings whose evaluation was contaminated by the very data being evaluated. The failure mode is the same at every scale: a system validating itself with information correlated to its own errors reports success it has not earned.

The strategic consequence deserves its own sentence, because much of the industry's current posture assumes the opposite. **Scaling one model does not buy verification.** A larger model is a stronger generator, and often a stronger critic of others, but its self-evaluations remain correlated with its own failure modes, and no parameter count changes that arithmetic. A bigger self cannot check itself.

4. The Answer: Independent Verification Across Multiple, Error-Independent Checkers

If self-corroboration is bounded by self-correlation, the remedy follows directly: corroboration must come from checkers whose errors are independent of the thing being checked. This is an old and rigorous idea. The Condorcet jury theorem shows that many independent, individually competent judges approach a reliability that no single judge attains; the same literature shows the guarantee collapses when the judges' errors are correlated, because, as the Stanford Encyclopedia of Philosophy's treatment puts it, "Common causes create correlations" (Dietrich & Spiekermann 2021; Ladha 1992). The result has been formally ported from voting theory to real judgment aggregation: crowd accuracy follows the same bias, variance, and covariance structure as machine ensembles, and the crowd's advantage shrinks as inter-judge correlation rises (Davis-Stober et al. 2014).

The critical nuance, and the one most often missed in current AI practice, is that **independence must be genuine, and genuine independence is hard to get**. The diagnostic question is not "how many checkers?" but "what do the checkers share?", because two checkers can look independent and fail as one. Across the monoculture literature the shared roots of correlation are recognizable, and four recur: shared training data (the same gaps, the same wrong cases); shared architecture (the same inductive biases, the same blind spots); a shared upstream foundation model (fine-tunes of one base are correlated by construction, and derivatives of a small number of base models are increasingly common); and a shared vendor or preprocessing pipeline (one organization's assumptions about what the data means, inherited by everything downstream). Checkers that share any of these roots are not independent judges; they are a common cause wearing two hats. Bommasani et al. (2022) found empirically that models sharing training data systematically fail on the same cases. Kleinberg & Raghavan (2021) showed where this leads at population scale: convergence by many decision-makers on a single algorithm reduces the overall quality of decisions across the population, even when that algorithm is individually the most accurate choice. "More models" is not the prescription. Differently wrong models are the prescription: different data, different architectures, different vendors, and ideally different modalities of checking altogether. An ecosystem that stacks near-clones of one foundation model and calls the stack an ensemble has reproduced the corroboration problem at industrial scale.

The self-referential version of the failure is now documented in the training loop itself. Shumailov et al. (2024) showed that models trained indiscriminately on model-generated content degrade, with the tails of the original distribution disappearing across generations. We cite this for the mechanism, not as an

inevitability: subsequent analyses, and a published author correction, make clear that collapse is avoidable when real data is retained and accumulated alongside synthetic data. Both halves of that finding support the same conclusion. Self-reference degrades; grounding in independent, real-world signal is what rescues it.

The same structure governs AI-evaluating-AI. LLM judges are often genuinely useful: Zheng et al. (2023) found that strong models reach high agreement with human evaluators, and the LLM-as-judge paradigm has real practical value. The same study also identified structural biases, position, verbosity, and notably self-enhancement, and Panickssery, Bowman & Feng (2024) demonstrated that when the generator and the judge are the same model, evaluation is biased toward the model's own outputs. The defensible claim is therefore two-sided and specific: AI evaluation is not worthless, but a model grading itself carries a structural, non-scaling bias in its own favor. Meanwhile, the intervention that reliably works points exactly where the theory says it should: across reasoning and planning benchmarks where self-critique fails, introducing an external, sound verifier into the loop restores performance (Stechly, Valmeekam & Kambhampati 2025).

What does genuine independence look like for Physical AI? Recall the three layers. The verification loop that matters is **Layer 2 checking Layer 3 against Layer 1**: models of the world testing what has been described or proposed against how the world actually behaves. Concretely, the adjudication layer draws on checkers whose error profiles are naturally uncorrelated. Agreement across checkers like these is informative exactly because they fail differently. Agreement among correlated estimators is false comfort. Independence is the active ingredient, and it is a property to be engineered and audited, not assumed.

5. "Trust, but Verify, Independently" as an Engineering Posture

It would be a fair challenge to ask whether this argument, however sound, is practical. The decisive answer is that it is already practice. In every mature engineering domain where the cost of being wrong is physical, independence of verification is not a slogan. It is codified, specified doctrine with a definition, a budget, and an institutional home.

The canonical statement is IEEE Std 1012, the verification and validation standard, which defines independent V&V by three properties: technical independence (the verifiers were not the developers and form their own understanding of the problem), managerial independence (the verification organization selects what to examine and reports findings without developer approval), and financial independence (the verification budget is not controlled by the development organization). NASA has operationalized exactly this model in a permanent, agency-level IV&V program, organizationally and budgetarily separated from the projects it examines. NASA's software engineering guidance states it directly: "The key parameters for independence are technical independence, managerial independence, and financial independence" (NASA SWE-141).

Just as telling is that the required degree of independence scales with the stakes. Under IEC 61508, the functional-safety standard for industry, the independence required of the safety assessor increases with the safety integrity level; at SIL 4, the highest level, the assessor must come from an independent organization entirely. ISO 26262 grades the independence of its confirmation measures by automotive safety integrity level. DO-178C, the software standard for airborne systems, requires that verification objectives at the highest design assurance levels be satisfied with independence: the verifier is not the author. The pattern

repeats beyond standards into regulation. The U.S. Nuclear Regulatory Commission does not simply audit an applicant's safety analysis; its staff performs independent confirmatory calculations with its own codes, and the statutorily independent Advisory Committee on Reactor Safeguards reviews both the application and the staff's evaluation. U.S. medical-device regulation requires that each design review include a reviewer without direct responsibility for the design stage under review (21 CFR 820.30(e)): even inside a single manufacturer, pure self-review of a design stage is forbidden by law.

Science itself runs on the same principle. The National Academies' consensus study on reproducibility states it plainly: "One of the pathways by which the scientific community confirms the validity of a new scientific discovery is by repeating the research that produced it" (NASEM 2019). Validity is conferred by independent repetition, not by the author's own certainty. And the doctrine is already migrating to AI: the NIST AI Risk Management Framework's MEASURE function contemplates independent review of AI systems, and ISO/IEC 42001 establishes third-party certification of AI management systems (NIST 2023; ISO/IEC 2023). These AI-specific instruments are emerging and voluntary, not settled mandate, and we present them as such. But their direction is unmistakable: the assurance pattern that every physical-consequence industry converged on is being written into AI governance.

Physical AI sits at the intersection of these traditions, and should inherit their posture. In engineering terms, that means: separate the generator from the verifier, structurally, not just procedurally. Treat error-profile diversity as a design requirement of the verification layer, procured and measured like any other requirement. Sequence verification before expensive or irreversible physical action, scaled to the stakes as the safety standards scale their independence requirements. And keep the checks legible. Here we name a tension honestly rather than wish it away: if the verification layer is itself an opaque learned system, it can undercut the very trust it exists to produce. Prediction is not understanding, and a surrogate that predicts is not a mechanism that explains. A verification layer for consequential domains should privilege auditable, interpretable checks, mechanistic models, physical constraints, empirical tests, documented human review, so that a regulator, a customer, or an engineer can see not only that a claim was checked but how.

6. Implications for Leaders Building or Buying Physical AI

For executives and technical leaders making Physical AI investments now, the argument compresses into four operating principles.

A bigger model is not a verification strategy. Scaling a single system, or ensembling near-copies of it, strengthens generation without manufacturing the independence that verification requires; the covariance term does not shrink (Ueda & Nakano 1996), self-correction without external signal does not reliably improve outcomes (Huang et al. 2024), and self-evaluated pipelines have produced overconfident results at literature scale (Kapoor & Narayanan 2023). If a roadmap's answer to "how will we know it is right?" is "a better model," the roadmap does not have an answer.

Evaluate vendors on independence, not accuracy alone. Accuracy claims describe the generator. Verification claims describe the checker, and the first question about any checker is what it shares with the thing it checks: training data, base model, vendor, incentive. Judges that share a foundation with the judged

carry structural self-preference (Panickssery et al. 2024), and monocultures fail together (Kleinberg & Raghavan 2021; Bommasani et al. 2022). Ask vendors to demonstrate error-profile diversity, not just benchmark scores.

Put the independent verification layer where the physical stakes are highest. The safety-critical canon offers a ready template: graded assurance, in which the required independence of the check scales with the consequence of the failure, as in IEC 61508 and ISO 26262. Not every AI output needs adjudication. Outputs that gate expensive experiments, physical designs, or control actions do, and the gate belongs before the irreversible step, not after.

Decide who owns the independent "no." This is ultimately an organizational question, and the hardest one, because an independent verdict that answers to the builder's schedule and budget is not independent. NASA's answer, technical, managerial, and financial independence, vested in a separately funded organization, is the proven template. Leaders should be able to name the function in their organization that is empowered to stop a physically consequential AI-driven action, and to say from whose budget it is paid. If the answer to either question is "the team building the system," the exam is still being graded by the student.

7. What Would Make This Argument Wrong

An argument that nothing should grade its own exam owes the reader the terms on which it would fail. We can name three.

First, the mathematics. If a method is demonstrated that strips shared, systematic error out of copies of a single model as dependably as genuine independence does, the covariance floor stops binding and the case for structurally separate checkers weakens to a preference. Decorrelation techniques already reduce the variance-type, random component of error, and we cited them approvingly above; what none has shown is the removal of the shared, systematic error that a common cause bequeaths to all of its copies. We know of no such method, but the claim is testable, and we state it that way deliberately.

Second, the direction of practice. We read safety standards, regulation, and procurement as converging on demanded independence. If over the coming years they do not, if assurance regimes at physical stakes settle for self-certification, then the direction we called unmistakable is not, whatever we think of the wisdom.

Third, the economics. Independent verification earns its place only where the failures it prevents cost more than the checking does. If in most real deployments that inequality runs the other way, the posture does not pay, and engineering postures that do not pay do not spread.

None of the three holds today, so far as we can see. That is why we hold the view, and it is how a reader would know if it stopped being right.

8. Conclusion

Physical AI is an old mission entering a new phase. High-performance computing has spent decades building Layer 2, models of the world, in service of Layer 1, the world itself. Language models have transformed Layer 3, our descriptions of the world. The convergence of these layers is the most

consequential development in advanced computing in a generation, and nothing in this paper argues against it. We argue for the discipline that makes it trustworthy: a verification loop in which models of the world check machine-generated claims against the world's actual behavior, and in which the checkers are genuinely independent of the thing being checked.

The core result is easy to state and hard to escape. Corroboration reduces error only between estimators that fail differently. A system reasoning under one set of weights shares its own blind spots, so its agreement with itself certifies confidence, not correctness. Independent verification is therefore structural, not optional, and everywhere civilization has already faced this problem, in avionics, nuclear power, medical devices, spaceflight, and science itself, it has converged on the same answer: separate the checker from the maker, and scale the separation with the stakes.

Nothing here rejects large language models, and nothing here rejects simulation. Independent verification is the discipline that connects them: it is how fluency at Layer 3 and predictive power at Layer 2 are converted into justified trust about Layer 1. That conversion is where NeoVerity works, and in our judgment it is where the next decade of high-integrity Physical AI will be won or lost.

You cannot grade your own exam. In the physical world, the cost of a wrong grade is not a wrong answer. It is a wrong outcome. Build the independent verification layer accordingly.

References

1. Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? Advances in Neural Information Processing Systems 35 (NeurIPS 2022). arXiv:2211.13972.
2. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. DOI 10.1023/A:1010933404324.
3. Choudhury, S., Czajka, J. J., Monteiro, L. M. O., et al. (2026). Networked intelligence: Active shared context graphs for human-AI team science. arXiv preprint arXiv:2607.13220 (preprint, not peer-reviewed). Pacific Northwest National Laboratory.
4. Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? Decision, 1(2), 79-101. DOI 10.1037/dec0000004.
5. Dietrich, F., & Spiekermann, K. (2021). Jury theorems. In The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/jury-theorems/
6. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning (ICML 2017). arXiv:1706.04599.
7. Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. ISBN 978-0-9825442-0-4.
8. Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. International Conference on Learning Representations (ICLR 2024). arXiv:2310.01798.
9. IEC 61508. Functional safety of electrical/electronic/programmable electronic safety-related systems. International Electrotechnical Commission (ed. 2.0, 2010).
10. IEEE Std 1012. IEEE Standard for System, Software, and Hardware Verification and Validation. IEEE.
11. ISO 26262:2018. Road vehicles: Functional safety. International Organization for Standardization.
12. ISO/IEC 42001:2023. Information technology: Artificial intelligence. Management system. ISO/IEC.

13. JCGM 100:2008. Evaluation of measurement data: Guide to the expression of uncertainty in measurement (GUM). Joint Committee for Guides in Metrology, BIPM. See also NIST/SEMATECH, e-Handbook of Statistical Methods, sec. 2.
14. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. DOI 10.1016/j.patter.2023.100804.
15. Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. Proceedings of the National Academy of Sciences, 118(22). DOI 10.1073/pnas.2018340118.
16. Krishnapriyan, A. S., Gholami, A., Zhe, S., Kirby, R. M., & Mahoney, M. W. (2021). Characterizing possible failure modes in physics-informed neural networks. Advances in Neural Information Processing Systems 34 (NeurIPS 2021). arXiv:2109.01050.
17. Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. Advances in Neural Information Processing Systems 7 (NIPS 1995), 231-238.
18. Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. American Journal of Political Science, 36(3), 617-634. DOI 10.2307/2111584.
19. LeCun, Y. (2022). A path towards autonomous machine intelligence (Version 0.9.2). OpenReview. Peer-reviewed companion: Journal of Statistical Mechanics (2024), DOI 10.1088/1742-5468/ad292b.
20. Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2021). Fourier neural operator for parametric partial differential equations. International Conference on Learning Representations (ICLR 2021). arXiv:2010.08895.
21. Lu, L., Jin, P., Pang, G., Zhang, Z., & Karniadakis, G. E. (2021). Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. Nature Machine Intelligence, 3, 218-229. DOI 10.1038/s42256-021-00302-5.
22. NASA. Software Engineering Handbook, SWE-141: Software Independent Verification and Validation. NASA. https://swehb.nasa.gov/
23. National Academies of Sciences, Engineering, and Medicine (NASEM). (2019). Reproducibility and Replicability in Science. The National Academies Press. DOI 10.17226/25303.
24. National Academies of Sciences, Engineering, and Medicine (NASEM). (2024). Foundational Research Gaps and Future Directions for Digital Twins. The National Academies Press. DOI 10.17226/26894.
25. National Transportation Safety Board (NTSB). (2019). Assumptions Used in the Safety Assessment Process and the Effects of Multiple Alerts and Indications on Pilot Performance. Safety Recommendation Report ASR-19-01. Accident nos. DCA19RA017 and DCA19RA101.
26. National Transportation Safety Board (NTSB). (2019). Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018. Highway Accident Report NTSB/HAR-19/03.
27. NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology.
28. Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. Advances in Neural Information Processing Systems 37 (NeurIPS 2024). arXiv:2404.13076.
29. Psaros, A. F., Meng, X., Zou, Z., Guo, L., & Karniadakis, G. E. (2023). Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. Journal of Computational Physics, 477, 111902. DOI 10.1016/j.jcp.2022.111902.
30. Rackauckas, C., Ma, Y., Martensen, J., et al. (2020). Universal differential equations for scientific machine learning. arXiv preprint arXiv:2001.04385 (preprint).
31. Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707. DOI 10.1016/j.jcp.2018.10.045.
32. RTCA DO-178C. (2011). Software Considerations in Airborne Systems and Equipment Certification. RTCA, Inc.

33. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631, 755-759. DOI 10.1038/s41586-024-07566-y. (Author correction published; see also counter-analyses, e.g., arXiv:2503.03150, showing collapse is avoidable when real data is accumulated.)
34. Stechly, K., Valmeekam, K., & Kambhampati, S. (2025). On the self-verification limitations of large language models on reasoning and planning tasks. International Conference on Learning Representations (ICLR 2025). arXiv:2402.08115.
35. Ueda, N., & Nakano, R. (1996). Generalization error of ensemble estimators. Proceedings of the IEEE International Conference on Neural Networks (ICNN'96), Vol. 1, 90-95. DOI 10.1109/ICNN.1996.548872.
36. U.S. Code of Federal Regulations. 21 CFR 820.30, Design controls; §820.30(e) requires each design review to include "an individual(s) who does not have direct responsibility for the design stage being reviewed." U.S. Food and Drug Administration. (This design-control requirement carries forward under the FDA Quality Management System Regulation, which incorporates ISO 13485:2016 by reference, effective February 2, 2026.)
37. U.S. Nuclear Regulatory Commission (NRC). Independent confirmatory calculations and design certification review practice, including independent review by the Advisory Committee on Reactor Safeguards (ACRS). https://www.nrc.gov/
38. Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181(8), 1065-1070. DOI 10.1001/jamainternmed.2021.2626.
39. Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track. arXiv:2306.05685.

About

NeoVerity Group

NeoVerity Group is a senior technology advisory practice working at the convergence of high-performance computing and artificial intelligence. It advises on AI strategy, HPC and AI infrastructure, government R&D programs, and technical due diligence, and publishes open research on high-integrity Physical AI. More about the practice.

Mike Vildibill

Founder and Principal of NeoVerity Group. Thirty-five years at the frontier of high-performance computing and AI, including senior product and general management leadership building the infrastructure behind large-scale scientific and AI systems. Get in touch.