Independent Verification for Physical AI
A Framework for Measuring Autonomous-System Performance Outside the Machine

Sean Kish
sean@paverly.com
www.paverly.com

Abstract. Physical AI systems increasingly make decisions with direct consequences in the physical world. Their performance, however, is still evaluated largely through telemetry, logs, incident reports, and safety claims generated by the systems and organizations being assessed. These sources may be accurate and useful, but they do not independently establish what occurred. This paper proposes a verification framework based on motion measured outside the autonomy stack. Every autonomous vehicle, robot, drone, or mobile machine ultimately expresses its decisions through physical movement. Motion therefore provides a universal, implementation-independent reference against which a system's reported operational state can be compared. Repeated reconciliation between independently observed motion and machine-reported state can produce auditable evidence, comparable performance measures, and increasing confidence over time. The approach complements rather than replaces simulation, testing, safety cases, standards, and internal telemetry by providing the independent physical reference those methods do not.

1. Introduction

Autonomous systems are moving from controlled demonstrations into transportation, logistics, manufacturing, agriculture, construction, defense, healthcare, and consumer applications. Software that once informed human action now directly controls how vehicles, robots, and machines behave in the physical world. As deployment expands, the question is no longer limited to whether an autonomous system functions correctly under selected conditions. Regulators, insurers, fleet operators, investors, commercial customers, and the public increasingly need to determine how reliably these systems perform in ordinary operation and whether the evidence supporting that performance can be independently evaluated.[1] [2]

Modern autonomous systems generate extensive records of their own operation, including position, velocity, localization quality, perception confidence, interventions, software health, actuator status, and safety events. These data are indispensable for development and operation, but they generally originate within the system or organization whose performance is being assessed. Manufacturers naturally possess the deepest knowledge of their technology, yet other parties must make consequential decisions without equivalent access. Insurers must assess operational risk,[3] regulators must evaluate public safety, fleet operators must compare platforms, and customers and investors must distinguish durable performance from claims that cannot be evaluated on a common basis.

The difficulty is not a lack of information. It is that company-specific measures such as autonomous miles, disengagements, collisions, interventions, and internally developed safety scores are often defined differently, collected under different conditions, and reported without common exposure data. NHTSA has cautioned that even federally collected crash data cannot be used directly to compare manufacturer safety performance when the relevant operational exposure is unavailable,[4] while researchers have similarly

questioned whether manufacturer-reported net-risk metrics can be compared or independently verified.[5] Additional disclosure can provide a more detailed account of what the system recorded, but it does not create a separate record of what happened in the physical world. A credible evaluation framework therefore requires a physical reference generated outside the system being evaluated.

2. The Self-Verification Problem

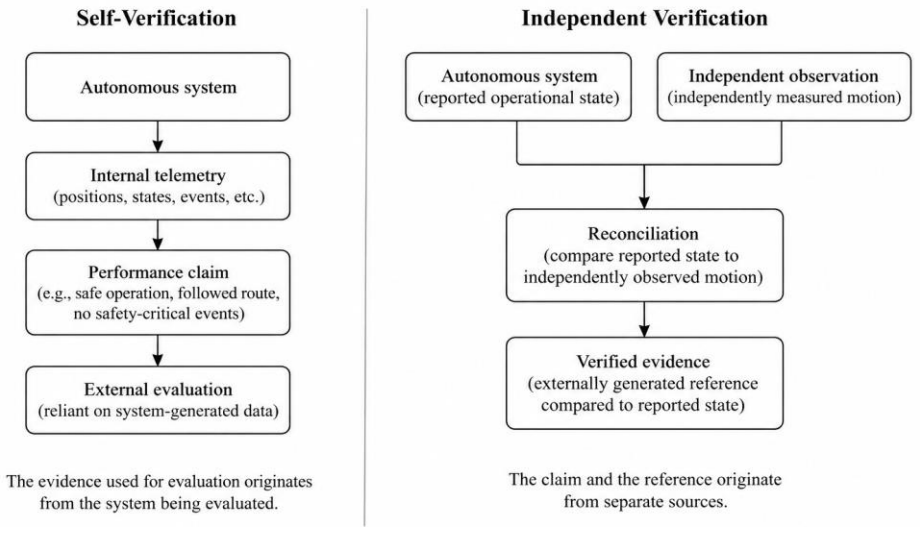
An autonomous system must continuously estimate its own state. It determines where it is, how it is moving, what it perceives, and whether its components are functioning correctly. This self-observation is essential to safe operation and is a hallmark of good engineering. It is not, however, independent verification. If a vehicle reports that its localization remained within specification throughout a journey, the report may be entirely correct. But when both the estimate and the evidence supporting it originate within the vehicle, an external evaluator still lacks an independent basis for confirming the result.

The same limitation applies when a system reports that it maintained a specified speed, followed an intended route, completed a fallback maneuver, or experienced no safety-critical event. Internal telemetry can support each claim, but the evidence remains dependent on the system's own sensors, software, timing, calibration, and reporting logic. This is not unique to autonomous systems: aircraft, medical devices, industrial equipment, and modern vehicles all rely on internal diagnostics. The limitation arises only when internally generated information becomes the principal evidence used by outside parties to evaluate the system as a whole. Recent NTSB findings have illustrated that manufacturer telematics designed for internal system evaluation may not be designed to satisfy external reporting or verification needs.[6]

Transparency and verification therefore serve different purposes. Transparency reveals what the system recorded, how it represented an event, and what its internal components reported. Independent verification compares that account with evidence produced separately from the system being evaluated. More disclosure may improve understanding, but it cannot remove the recursive problem created when the machine both produces the behavior and supplies the principal evidence used to assess it. A separate reference is required.

A similar principle underlies zero-trust architecture in cybersecurity, which does not assume that a user, device, or system should be trusted solely because it has previously been authorized or operates within a defined boundary. Instead, trust is continually evaluated through explicit verification. Applied to Physical AI, the analogy is not that machine-generated telemetry should be rejected, but that it should not serve as the sole basis for external confidence. Reported operational state becomes more credible when it is repeatedly compared with evidence generated independently from the system being assessed.[7]

Figure 1. The Self-Verification Problem and Independent Verification



3. Requirements for Independent Verification

An independent verification framework must do more than collect additional data. It must produce evidence that remains useful across manufacturers, software architectures, operating environments, and applications. The first requirement is operational independence: the evidence should originate outside the autonomy stack being evaluated and should not depend entirely on the same sensors, timing sources, calibration processes, software pipelines, or reporting logic used by the machine to estimate its own state. Independence is not absolute, and some shared infrastructure may be unavoidable, but the relevant dependencies should be identified and minimized so that a common failure cannot invalidate both the reported state and the reference used to assess it.

The second requirement is implementation neutrality. Autonomous systems use different combinations of cameras, radar, lidar, GNSS, maps, learned models, planners, and control architectures. A verification method should not require access to any particular design, nor should it require manufacturers to disclose source code, training data, perception models, or other competitively sensitive material. Its purpose is to verify observable performance, not to reproduce the system's internal reasoning. This distinction is consistent with existing safety-assurance approaches, including claim-based safety cases, which organize evidence without requiring every evaluator to reconstruct the underlying system.[8]

The third requirement is comparability. Measurements, event definitions, uncertainty estimates, and exposure units must be sufficiently consistent for evidence to be interpreted across systems, fleets, software releases, and time. Comparability does not mean treating unlike operating domains as equivalent. It means applying common methods while preserving the conditions under which each observation occurred. The fourth requirement is scalability. A framework based only on occasional inspections, closed-course tests, or post-incident investigations cannot provide a persistent view of deployed performance. Verification must be economical enough to operate across large numbers of machines and operating hours.

Finally, confidence must accumulate through repeated observation. A single test, route, or incident cannot characterize an autonomous system. Evidence becomes more useful as it spans operating conditions, software versions, platforms, and time, with uncertainty and data quality retained rather than hidden.[9] Taken together, these requirements describe a framework that is operationally independent, implementation-neutral, protective of proprietary information, comparable, scalable, and longitudinal. They establish the criteria against which any proposed independent verification technology should be evaluated.

4. Physical Behavior as the Common Reference

Independent verification requires a reference that exists outside the autonomy stack and remains meaningful across systems that differ substantially in design. Autonomous vehicles, trucks, delivery robots, drones, industrial machines, and humanoids may use entirely different sensing and control architectures, but all of them ultimately express their decisions through physical behavior. The internal path from perception to planning to control may be proprietary and platform-specific; the resulting movement occurs in a shared physical world.

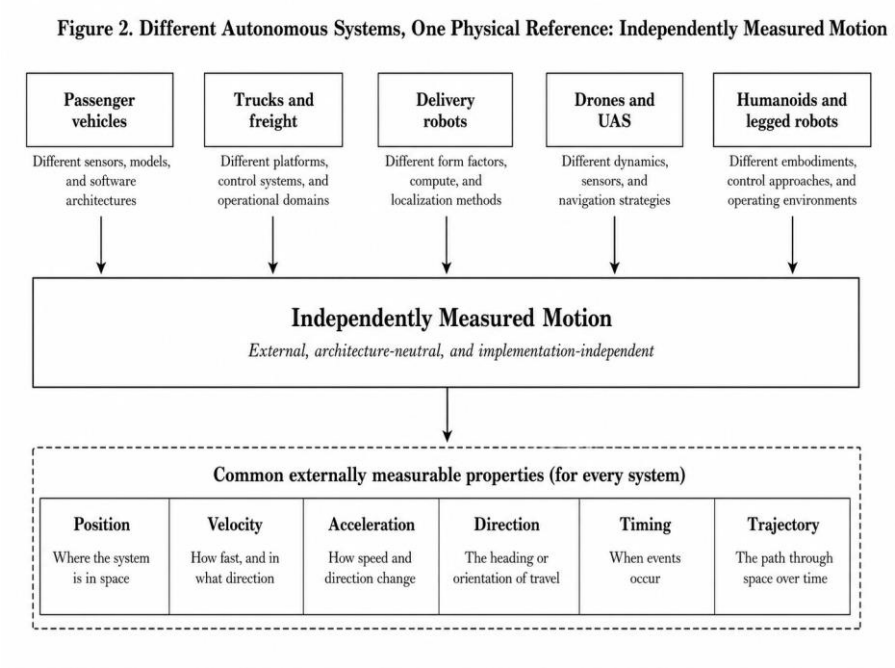
Motion is especially suitable as a common reference because every mobile autonomous system must determine where it is, how it is moving, and how that movement changes over time. Position, velocity, acceleration, direction, timing, and trajectory are not manufacturer-specific concepts. They are observable properties of physical systems and can be measured without reproducing the machine's perception, planning, or control architecture. This does not make motion a complete description of autonomy, but it makes motion a practical basis for comparing the machine's reported operational state with an external record of what it physically did.

Motion also provides continuity that event-based measures do not. Collisions, disengagements, interventions, and fallback events are discrete and relatively infrequent. Motion exists before, during, and after those events and throughout ordinary operation. Continuous observation can therefore identify both persistent agreement and material divergence between reported and observed state, rather than limiting evaluation to rare failures. This is important because performance cannot be inferred reliably from incident counts alone, particularly when exposure and operating conditions differ.[10]

A physical reference does not need to explain why every decision was made. An external velocity measurement, for example, cannot determine whether acceleration was appropriate, but it can establish whether the vehicle's reported velocity and acceleration corresponded to its observed movement. Context remains necessary to interpret behavior: road geometry, traffic, weather, nearby objects, operating rules, and the system's intended function may determine whether a maneuver was reasonable. The role of independent motion measurement is narrower and more defensible. Internal telemetry records what the system believed and reported; external measurement records what the system physically did. Agreement between those records supports the reported state, while a material difference creates a defined event requiring explanation.

Different autonomous systems compute behavior differently, but all mobile systems express behavior through measurable motion.

Figure 2. Different Autonomous Systems, One Physical Reference: Independently Measured Motion



5. Independent Motion Measurement

A common physical reference is useful only if it is measured with sufficient independence from the autonomy stack. The measurement system should therefore be separate from the sensors, software, and reporting logic used by the machine to estimate its own state, while any remaining shared dependencies should be identified and bounded. Its function is deliberately limited. It does not need to perceive surrounding objects, identify hazards, reconstruct the machine's reasoning, or control its behavior. Those responsibilities remain within the autonomous system. The external measurement system establishes a second record of motion against which the machine's reported operational state can be compared.

Two measurements are not independent merely because they are produced by different components. If both rely on the same localization source, clock, calibration process, map, communications channel, or software pipeline, they may reproduce the same error. A credible reference should therefore minimize correlated failure modes and characterize the dependencies that cannot be eliminated. The relevant measurements may include position, velocity, acceleration, direction, timing, and trajectory. They need not all be produced by one instrument, and no single measurement method will be appropriate for every platform or operating environment. The requirement is that the resulting reference be sufficiently independent, accurate, synchronized, and persistent for the intended comparison.

One implementation of this function is an independent motion-measurement system designed to observe host-platform movement separately from the machine's primary autonomy stack. Such a system does not determine whether the machine selected the correct action, nor does it replace perception, planning, or control. Its role is narrower: to produce an external record of what the platform physically did. That

limitation is important because it keeps the verification claim testable. A system designed to observe motion does not need to reproduce the entire autonomy problem in order to verify reported motion. Its output is not yet a performance assessment; it is an independently produced reference record that becomes useful through reconciliation.

6. Reconciliation

Reconciliation compares the operational state reported by the machine with the state inferred from independent observation. Let the machine-reported state at time t be represented by a vector R(t), and let the independently observed state be represented by O(t):

$$R(t) = [R_1(t), R_2(t), \ldots, R_n(t)]^T$$

$$O(t) = [O_1(t), O_2(t), \ldots, O_n(t)]^T$$

The components may represent corresponding quantities such as position, velocity, acceleration, heading, or timing, after conversion to a common coordinate frame, unit system, and time base. The difference between the reported and observed states is the error vector:

$$e(t) = R(t) - O(t)$$

For each relevant component i, the observed difference is:

$$e_i(t) = R_i(t) - O_i(t)$$

A nonzero difference does not by itself indicate a failure. Every measurement has uncertainty, and two systems may sample, filter, timestamp, or represent the same movement differently. Verification therefore requires defined tolerances rather than exact numerical identity. If $\tau_i(t)$ represents the allowable difference for component i under the applicable conditions, the reported and observed values are treated as consistent when:

$$|e_i(t)| \leq \tau_i(t)$$

This component-level comparison is preferable to a single undifferentiated error value because position, velocity, acceleration, heading, and timing have different units, uncertainty characteristics, and operational significance. Combining them into one metric would require an explicit normalization and weighting method. The tolerance for each component should account for the uncertainty of both measurement sources, installation and calibration effects, synchronization quality, platform dynamics, environmental conditions, and the purpose of the assessment. Established metrological practice similarly treats a measurement result as incomplete unless its associated uncertainty is stated.[11]

Time alignment is a necessary part of the comparison. The reported and observed states must refer to the same physical instant, not merely carry similar timestamps. If the machine's reporting pipeline introduces a latency $\delta(t)$, the relevant comparison may be between R(t) and $O(t - \delta(t))$, or between two records after estimating and correcting the offset. Clock drift, processing delay, communications latency, and filtering can otherwise create an apparent discrepancy even when both systems measured the same motion correctly. Angular quantities also require appropriate treatment: heading differences, for example, should be calculated as circular differences so that values near zero and 360 degrees are not treated as widely separated.

A validation envelope provides a structure for applying these comparisons across multiple variables, times, and operating conditions. It is not a single fixed threshold. It defines the bounded relationships within which reported and independently observed motion are treated as consistent for a specified purpose. Context remains necessary because consistency does not establish that a maneuver was appropriate. A rapid deceleration may reflect unstable control or a correct response to an obstacle. Reconciliation answers the narrower question of whether the machine's account corresponds to independently observed behavior; environmental and operational evidence is then used to interpret what that behavior means.

When the reported and observed states remain within the applicable envelope, the evidence supports the machine's account of its motion. When they diverge, the discrepancy becomes a defined event that can be examined, classified, and retained. The framework should not presume that every discrepancy originates with the autonomous system. The external instrument, its installation, calibration,

synchronization, data transmission, or contextual record may also be responsible. Reconciliation therefore changes the evidentiary status of machine-generated telemetry without declaring either source infallible. Before comparison, telemetry is a report produced by the system. After comparison with a sufficiently independent reference, the corresponding portion can be treated as externally supported evidence.

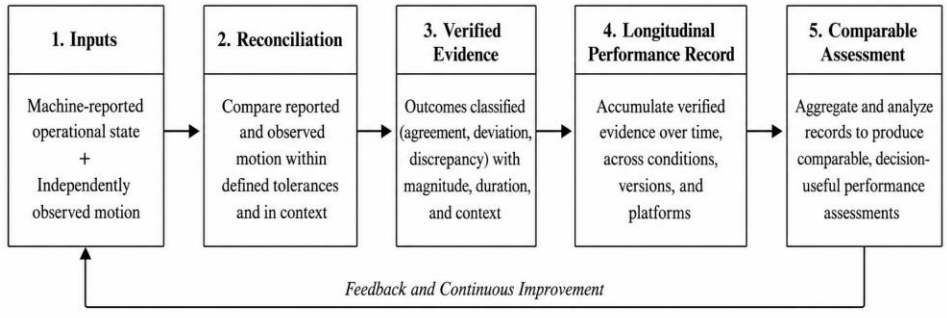
7. Evidence Accumulation

A single successful comparison provides little basis for characterizing an autonomous system. Operating environments are variable, software changes frequently, and serious failures are generally too rare for isolated events to support broad conclusions. Independent verification should therefore be treated as an evidence-accumulation process in which each reconciled observation contributes to a longitudinal record of where, when, and under what conditions reported state agreed with independently observed motion. Confidence develops through repeated correspondence across operations rather than through a one-time pass-or-fail judgment.[12]

The resulting record should preserve more than a binary indication of agreement. It should retain the magnitude and duration of discrepancies, the variables involved, applicable tolerances, measurement uncertainty, synchronization quality, operating context, platform and software version, and the disposition of any anomaly. This allows later analysis to distinguish transient noise, measurement faults, and isolated events from persistent or systematic differences. The evidentiary value also depends on coverage. Agreement observed only at low speed in clear conditions does not establish performance in rain, dense traffic, construction zones, degraded localization environments, or other unobserved conditions. Conclusions should remain bounded by the operating domains, behaviors, and versions actually represented in the record.

The appropriate unit of analysis will depend on the question being asked. An engineer may investigate millisecond-level differences during a maneuver, while a fleet operator may compare discrepancy rates across vehicles, sites, or releases. An insurer may require exposure-normalized trends over months, and a regulator may examine whether performance remains within a defined operating domain. The same verified observations can support these different analyses if their provenance, uncertainty, context, and limitations are retained. Accumulated evidence does not eliminate uncertainty; it makes uncertainty measurable and reveals where confidence is strong, where it remains limited, and where additional observation is required.

Figure 3. From Reports to Comparable Performance: The Verification Architecture



8. From Verified Evidence to Comparable Performance

Verified observations become more useful when they can be interpreted consistently across systems, fleets, operating domains, and software versions. The purpose of comparison is not to erase meaningful differences among autonomous systems, but to evaluate them through common definitions, disclosed methods, and evidence of known quality. A comparable assessment therefore requires normalization for exposure, context, uncertainty, and data completeness. Events must be classified consistently, statistical

confidence must be distinguished from raw counts, and the methodology must state where comparison is valid and where the available evidence is insufficient.

Two systems should not be compared solely on the number of observed discrepancies if one has operated longer, encountered more difficult conditions, or produced more complete contextual data. Exposure may need to be expressed through distance, operating hours, tasks completed, maneuvers performed, or another domain-appropriate unit. Operating conditions must also be preserved because performance in a narrow, controlled domain is not equivalent to performance across a broader or more complex environment. Missing records, uncertain synchronization, incomplete context, or inadequate exposure should reduce confidence and may prevent a comparative assessment from being issued. These constraints are consistent with broader concerns about comparing autonomous-system performance without common definitions and exposure measures.[13]

A comparative assessment framework can convert independently verified evidence into standardized performance measures across manufacturers, fleets, software generations, and operating conditions. The resulting score or rating is not the evidence itself; it is a summarized interpretation of the underlying record. Any assessment should therefore remain traceable to its inputs, methods, assumptions, operating-domain coverage, and confidence level. Users should be able to distinguish among an early directional assessment, a statistically supported comparison within a defined domain, and a mature rating based on broad operational evidence. A single numerical output without those distinctions would imply more certainty than the record supports.

The framework should also preserve the separation between measurement and judgment. Independent motion measurement establishes what occurred. Reconciliation determines whether the machine-reported and independently observed states were consistent. Contextual analysis helps interpret the behavior, and scoring summarizes selected aspects of the verified record for a defined purpose. The same underlying evidence may therefore support different uses: engineers may examine discrepancy patterns, fleet operators may compare releases or operating sites, insurers may evaluate exposure-normalized risk indicators, and regulators may assess performance within defined operating domains. Comparable performance depends not only on independent measurement, but also on disciplined limits regarding what is being compared, under which conditions, using what exposure, and with what confidence.

9. Governance

A common measurement and scoring framework will affect manufacturers, operators, insurers, regulators, customers, and the public. Its credibility therefore depends not only on the technical performance of the measurement system, but also on how the rules are created, revised, applied, and challenged. The organization operating the technical infrastructure may collect data, perform reconciliation, maintain systems, and produce assessments, but it should not possess unrestricted authority to define every scoring criterion, weighting, threshold, or disclosure rule without external participation.

A multi-stakeholder industry body could provide that governance function. Its role would be to establish and revise methodologies, taxonomies, confidence requirements, disclosure practices, and review procedures with participation from manufacturers, operators, insurers, regulators, technical experts, researchers, and other affected parties. This structure would not eliminate disagreement, nor should it. Different stakeholders will have different objectives and tolerances for risk. Its purpose is to make those differences visible and subject to a defined process rather than leaving methodology under unilateral control.

Governance should include published criteria, documented revisions, conflict-of-interest rules, technical review, and procedures for challenging, correcting, or withdrawing an assessment. Material methodological changes should be versioned so that historical results remain interpretable. This distinction between infrastructure and governance is important: the technical system produces and reconciles evidence, while the governance system determines how that evidence may be interpreted for common use. Comparable ratings in other industries have generally depended on both repeatable methods and institutional credibility, rather than on measurement technology alone.[14]

10. Relationship to Existing Approaches

Independent verification does not replace the methods already used to develop and evaluate autonomous systems. Simulation allows developers to test large numbers of scenarios before deployment; closed-course testing provides controlled evaluation of selected behaviors; safety cases organize evidence supporting claims about acceptable risk; standards and certification define processes, requirements, and technical expectations; internal telemetry and diagnostics support operation and investigation; and incident analysis examines failures after they occur. Each remains necessary, but none by itself creates a persistent external record of deployed physical behavior across different systems.

The limitation arises from the function of each method. Simulation depends on models, assumptions, and scenarios selected by the developer or evaluator. Testing examines bounded conditions. Safety cases depend on the quality, completeness, and relevance of the evidence submitted. Standards define requirements but do not necessarily provide continuous operational measurement. Internal telemetry remains generated by the system being evaluated, while incident investigation is generally episodic and retrospective. Existing safety-assurance frameworks, including UL 4600, appropriately emphasize structured evidence and argument, but they do not eliminate the need for independently generated operational evidence.[15]

Independent motion verification complements these methods by adding a separate physical reference during real-world operation. It can support a safety case, test whether internal telemetry corresponds to observed behavior, provide evidence across software releases, and identify events requiring further analysis. The framework should not be described as a complete safety system: it does not perceive every hazard, evaluate every decision, or replace engineering judgment. It verifies a defined class of observable claims and contributes evidence to a broader evaluation process.

11. Limitations

The proposed approach has important limitations. Motion is not equivalent to safety: a system can move exactly as reported and still make a poor decision. Independent motion measurement verifies observable state, not whether the underlying decision was correct. Interpretation also requires context, including road geometry, weather, traffic, nearby objects, operating rules, task requirements, and system intent. Contextual data may contain its own uncertainty and may depend on sources that are not fully independent.

The external measurement system also introduces error. Accuracy, calibration, installation, synchronization, environmental sensitivity, data loss, and hardware failure must be characterized, and independence should not be mistaken for correctness. Shared dependencies may further weaken the reference if the external system relies on the same timing source, map, GNSS correction, or communications channel as the machine being evaluated. The degree of independence should therefore be documented and tested rather than asserted.

Comparison also requires adequate and representative exposure. Sparse data may identify discrepancies but cannot reliably characterize overall performance, particularly across operating domains that have not been observed. Confidence should reflect the quantity, diversity, and representativeness of the evidence. Scoring introduces additional judgment because choices about weighting, normalization, thresholds, and aggregation can materially affect the result. Those choices must be disclosed, governed, and evaluated against observed outcomes. Finally, the technical accuracy of the measurement, the usefulness of reconciliation, and the predictive value of any resulting score must be demonstrated across multiple platforms and operating conditions. These limitations define the conditions under which independent verification can be used credibly rather than weakening the case for it.[16]

12. Conclusion

Physical AI systems increasingly act without direct human control, yet much of the evidence used to evaluate them is produced by the systems and organizations being assessed. Internal telemetry, simulation, testing, safety cases, standards, and diagnostics remain essential, but they do not provide a fully independent record of real-world behavior. A separate physical reference is therefore required if external

stakeholders are to distinguish between a system's account of its operation and independently supported evidence of what occurred.

Motion is suitable for this role because it is universal across mobile autonomous systems, observable outside the autonomy stack, and measurable over time. Comparing independently observed motion with machine-reported operational state can convert portions of self-reported telemetry into externally supported evidence. The comparison must account for uncertainty, synchronization, context, platform-specific tolerances, and the limits of the available exposure. Its value increases through repeated observation across operating conditions, software versions, and fleets.

Independent verification does not replace the systems used to design, test, certify, or operate autonomous machines. It adds an external reference through which their reported behavior can be examined, accumulated, and compared. As Physical AI becomes more widely deployed, such verification may become a standard component of how autonomous-system performance is evaluated and governed.

Endnotes

1. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (January 2023). The framework emphasizes valid and reliable measurement, ongoing monitoring, and evidence of performance to support trust across developers, deployers, and other stakeholders in AI systems. Available at: https://doi.org/10.6028/NIST.AI.100-1

2. Consumer Reports, Comments to NHTSA on Incident Reporting for Automated Driving Systems and Level 2 Advanced Driver Assistance Systems (2026). Consumer Reports cites a nationally representative survey in which 89 percent of respondents supported requiring automated-vehicle companies to report safety-relevant incidents to regulators. Available at: https://advocacy.consumerreports.org/research/consumer-reports-comments-to-nhtsa-on-incident-reporting-for-automated-driving-systems-and-level-2-adas/

3. S&P Global Mobility, "How Autonomous Vehicles Will Change the Future of Car Insurance," September 3, 2025. The analysis describes how increasing automation changes underwriting, liability, claims investigation, and insurers' need for vehicle-behavior and software-version data. Available at: https://www.mobilityglobal.com/en-us/automotive-insights/blog/autonomous-vehicles-future-of-car-insurance

4. National Highway Traffic Safety Administration, "NHTSA Releases Initial Data on Safety Performance of Advanced Vehicle Technologies," June 15, 2022. NHTSA states that its crash data are not normalized by deployed vehicles or vehicle miles traveled and therefore cannot be used to compare manufacturers' safety performance. Available at: https://www.nhtsa.gov/press-releases/initial-data-release-advanced-vehicle-technologies

5. Philip Koopman and William H. Widen, "Breaking the Tyranny of Net Risk Metrics for Automated Vehicle Safety," Safety-Critical Systems eJournal (2024). The authors explain the difficulty of establishing comparable human-driver baselines and argue that aggregate net-risk measures are insufficient, particularly during early deployment. Available at: https://users.ece.cmu.edu/~koopman/pubs/Koopman2024_TyrannyNetRiskMetrics.pdf

6. National Transportation Safety Board, Fatal Crashes Between Vehicles Operating in Hands-Free Partial Automation Mode and Stationary Vehicles in San Antonio, Texas, and Philadelphia, Pennsylvania, Highway Investigation Report NTSB/HIR-26/02 (2026). The NTSB found that manufacturer telematics programs developed for internal system evaluation were not designed to capture all crash-related information required for external reporting and that the resulting federal database was likely incomplete. Available at: https://www.ntsb.gov/investigations/AccidentReports/Reports/HIR2602.pdf

7. Scott Rose, Oliver Borchert, Stu Mitchell, and Sean Connelly, Zero Trust Architecture, National Institute of Standards and Technology, Special Publication 800-207 (August 2020). The publication describes zero trust as eliminating implicit trust based on network location or prior authorization and requiring access decisions to be continually evaluated using explicit verification. The analogy here is limited to the principle of independent, repeated verification; it does not imply that Physical AI verification and cybersecurity architecture are technically equivalent. Available at: https://doi.org/10.6028/NIST.SP.800-207

8. ANSI/UL 4600, Standard for Safety for the Evaluation of Autonomous Products. UL 4600 uses a claim-based approach to safety-case construction and addresses risk analysis, testing, validation, data integrity, lifecycle concerns, metrics, and conformance assessment while remaining technology-neutral. It does not establish universal performance benchmarks or continuous operational ratings. Overview available at: https://oecd.ai/en/catalogue/tools/ansiul-4600-standard-for-safety-for-the-evaluation-of-autonomous-products

9. Marjory S. Blumenthal, Laura Fraade-Blanar, Ryan Best, and James M. Anderson, Measuring Automated Vehicle Safety: Forging a Framework, RAND Corporation, RR-2662-RC (2018). The report proposes a structured framework for measuring

automated-vehicle safety while emphasizing operating context, common terminology, information sharing, evolving software, and the limitations created by insufficient real-world exposure. Available at: https://www.rand.org/pubs/research_reports/RR2662.html

10. Ibid. RAND notes that statistically meaningful comparisons may require hundreds of millions of miles or more, that events occurring before sufficient exposure is accumulated should be treated as case studies, and that common taxonomies are needed to preserve operational context.

11. Joint Committee for Guides in Metrology, Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement, JCGM 100:2008. The guide establishes that a measurement result should be accompanied by an evaluation of its uncertainty and provides the internationally accepted framework for expressing that uncertainty. Available at: https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf

12. Francesca M. Favaro, Trent Victor, Henning Hohnhold, and Scott Schnelle, "Interpreting Safety Outcomes: Waymo's Performance Evaluation in the Context of a Broader Determination of Safety Readiness" (2023). The authors describe safety evaluation as a process of continuous confidence growth through in-use monitoring and emphasize combining aggregate outcomes with event-level analysis and other forms of evidence. Available at: https://arxiv.org/abs/2306.14923

13. Blumenthal et al., Measuring Automated Vehicle Safety: Forging a Framework. The report emphasizes the need for common definitions, operating-domain context, exposure information, and methods that remain meaningful as systems change through frequent software updates. Available at: https://www.rand.org/pubs/research_reports/RR2662.html

14. International Organization of Securities Commissions, Code of Conduct Fundamentals for Credit Rating Agencies, revised March 2015; see also ISO/IEC 17065:2012, Conformity Assessment—Requirements for Bodies Certifying Products, Processes and Services. These frameworks illustrate that credible third-party assessments depend not only on technical methodologies, but also on transparency, independence, conflict controls, documented procedures, review, and institutional governance.

15. ANSI/UL 4600, Standard for Safety for the Evaluation of Autonomous Products. The standard provides a structured safety-case framework but does not define general pass/fail performance criteria, acceptable risk levels, or comparative road-testing benchmarks. Independent operational measurement would therefore complement rather than replace the safety-case function.

16. National Institute of Standards and Technology, "Artificial Intelligence Test, Evaluation, Validation and Verification (TEVV)." NIST describes reliable measurement and evaluation as essential to trustworthy AI and emphasizes evaluation of technologies in use, identification of technical gaps and limitations, and development of metrics and methods appropriate to emerging applications. Available at: https://www.nist.gov/ai-test-evaluation-validation-and-verification-tevv