SA SENTINEL AUTHORITY

WHITE PAPER

Process Attestation vs. Behavioral Attestation

Why Governing AI Systems Is Not the Same as Verifying Their Behavior

February 2026 — Public Document

Version 1.0 — February 2026 — Public Document

Sentinel Authority
sentinelauthority.org

CONTENTS

Executive Summary

Every major AI governance framework in use today—ISO/IEC 42001, NIST AI RMF, the EU AI Act, SOC 2—verifies that an organization has the right processes in place. None of them verify that the AI system actually behaves as those processes require. ODDC closes this gap by certifying behavioral conformance: not what you say your system does, but what your system demonstrably does under sustained autonomous operation.

This white paper provides a detailed comparative analysis of the world's leading AI governance and safety frameworks, examining each one's specific provisions for verifying autonomous system behavior at runtime. It demonstrates, framework by framework, that the current standards ecosystem addresses organizational governance comprehensively while leaving a critical gap in behavioral verification—the gap between documented intent and operational reality.

The consequences of this gap are not theoretical. ECRI, the independent healthcare research organization, ranked AI systems deployed without proper oversight as the number one health technology hazard for 2025. The Department of Justice has subpoenaed pharma and digital health companies over AI deployed in electronic medical record systems. A 2024 academic analysis identified 51 court cases involving software-related patient injuries from clinical decision support, drug management, and surgical robotics. In autonomous transportation, repeated software recalls affecting thousands of vehicles demonstrate that governance processes did not prevent systematic behavioral failures.

This paper argues that the current standards are not wrong—they are incomplete. ODDC is not a replacement for ISO 42001 or the NIST AI RMF. It is the missing behavioral layer that makes these governance frameworks enforceable at the system level.

1. The Governance Achievement

Before examining what is missing, it is important to acknowledge what has been accomplished. The AI governance ecosystem that has emerged since 2023 represents a significant advance in how organizations manage AI risk. NIST's AI Risk Management Framework provides a structured approach to identifying, assessing, and mitigating AI risks. ISO/IEC 42001 provides the world's first certifiable AI management system standard. The EU AI Act establishes the world's first comprehensive AI regulation with binding obligations. These frameworks, individually and collectively, have established essential organizational infrastructure for AI governance.

But there is a consistent pattern across all of them: they verify what an organization does to manage AI systems. They do not verify what the AI systems do. This is not an oversight—it reflects a deliberate design choice. Management system standards have always focused on organizational processes because these are what auditors can evaluate using established conformity assessment methodologies. The problem is that for autonomous systems operating in safety-critical domains, organizational processes are necessary but not sufficient. A hospital can have world-class AI governance policies and still deploy a clinical decision support system that makes dangerous recommendations. A trucking company can maintain comprehensive safety management documentation and still operate autonomous vehicles with software defects that affect the entire fleet.

2. Framework-by-Framework Analysis

The following analysis examines each major governance framework's specific provisions for verifying autonomous system behavior at runtime. For each framework, we identify what it verifies, what it does not verify, and where the behavioral gap exists.

2.1 ISO/IEC 42001: AI Management Systems

ISO/IEC 42001, published in December 2023, is the world's first international standard specifically designed for AI management systems. It follows the Plan-Do-Check-Act methodology familiar from ISO 9001 and ISO 27001, providing a structured framework for establishing policies, assessing risks, implementing controls, and conducting management reviews. As of 2025, a CSA benchmark report indicates that 76% of organizations plan to pursue frameworks like ISO 42001.

What ISO 42001 verifies: That the organization has established an AI management system with defined policies, roles, and responsibilities. That risk assessments have been conducted covering the AI system lifecycle. That documentation exists for AI system design, development, and deployment. That management reviews occur on a regular cadence. That processes exist for monitoring, measurement, and continual improvement.

What ISO 42001 does not verify: That the AI system actually operates within the boundaries defined by those policies. That the risk controls documented in the management system are enforced at runtime. That the system's behavior matches its documented specifications under sustained autonomous operation. ISO 42001 explicitly focuses on organizational governance rather than specific technical implementations. A UNIDO analysis of the standard noted that there is currently no ecosystem of conformity assessment for digital services that is equivalent to that of tangible or manufactured products, and warned that regulators inserting product-certification requirements without understanding this gap risk creating requirements that are impossible or extremely costly to fulfill.

ISO/IEC 42001 certifies that your organization governs AI responsibly. It does not certify that your AI system behaves responsibly. An organization can hold full ISO 42001 certification while deploying systems that violate the very policies the certification attests to.

2.2 NIST AI Risk Management Framework (AI RMF 1.0)

The NIST AI RMF, released in January 2023 with expanded companion playbooks through 2024–2025, provides a voluntary, technology-agnostic framework organized around four core functions: Govern, Map, Measure, and Manage. It has become one of the most influential AI governance frameworks globally, particularly for U.S. organizations. NIST emphasizes outcome-focused practices and iterative improvement, explicitly avoiding prescriptive technical requirements.

What the AI RMF verifies: That governance structures exist with defined roles, policies, and escalation paths. That AI risks have been identified, categorized, and mapped to organizational context. That measurement approaches exist for evaluating AI system trustworthiness across dimensions including fairness, security, and transparency. That risk management processes are documented and continuously improved.

What the AI RMF does not verify: That the measurement approaches are applied to runtime behavior. That governance processes translate into enforceable operational constraints. That an autonomous system's behavior remains within the risk tolerances defined during the Map and Measure phases. NIST's own standardization plan, NIST AI 100-5e2025, explicitly acknowledges that conformity assessment with other standards is a Tier 2 priority item that requires more scientific work or maturity before standardization. The plan states that facilitating implementation of AI standards may require creating verification and validation tools and conformity assessment procedures—tools that do not yet exist in standardized form.

2.3 EU AI Act (Regulation 2024/1689)

The EU AI Act entered into force on August 1, 2024, establishing the world's first comprehensive, binding AI regulation. It creates a risk-based classification system with specific obligations for high-risk AI systems, including those used in transportation, critical infrastructure, healthcare, and other safety-critical domains. High-risk system obligations become applicable on August 2, 2026.

What the AI Act requires: Conformity assessments for high-risk AI systems, including quality management systems, technical documentation, record-keeping, transparency, human oversight, and accuracy, robustness, and cybersecurity requirements. The Act mandates that providers ensure their AI systems achieve an appropriate level of accuracy, robustness, and cybersecurity throughout their lifecycle.

Where the behavioral gap exists: The Act requires conformity assessments but the standards needed to perform them are not ready. CEN and CENELEC, the European standards organizations tasked with developing harmonized standards, were unable to meet the August 2025 deadline. The first harmonized AI standard—prEN 18286 for AI quality management systems—only entered public enquiry on October 30, 2025. In November 2025, the European Commission proposed extending the timeline, linking high-risk compliance deadlines to standards availability. The GPAI Code of Practice was released on July 10, 2025, but it addresses transparency and governance obligations for general-purpose AI, not runtime behavioral verification.

The critical point is that the EU AI Act creates a regulatory mandate for conformity evidence that does not yet exist. Providers will need to demonstrate that their systems meet accuracy, robustness, and cybersecurity requirements—but no harmonized standard defines how to measure or verify these properties for autonomous systems at runtime. ISACA has noted that ISO 42001 could be a fast track to readiness but will never be a legal shield—it builds governance infrastructure but does not satisfy the Act's technical compliance requirements.

2.4 SOC 2 and Related Assurance Frameworks

SOC 2 (Service Organization Control Type 2), developed by the AICPA, provides assurance over five trust service criteria: security, availability, processing integrity, confidentiality, and privacy. It is widely used in the technology sector and increasingly applied to AI-powered services. SOC 2 audits evaluate the design and operating effectiveness of controls over a period of time (typically 6–12 months).

What SOC 2 verifies: That security controls are in place and operating effectively. That data processing maintains integrity according to defined criteria. That system availability meets commitments. That confidentiality and privacy controls function as designed.

What SOC 2 does not verify: That an AI system's outputs are correct or safe. That the system operates within its designed operational domain. That autonomous decision-making adheres to specified constraints. SOC 2 was designed for service organizations processing data, not for autonomous systems making consequential decisions in physical environments. Applying SOC 2 to autonomous systems creates an assurance gap: the audit confirms that controls exist around the system's infrastructure without evaluating whether the system's behavior is constrained.

2.5 Sector-Specific Frameworks

Several sector-specific frameworks have attempted to address AI safety within their domains, each with the same structural limitation:

FDA Software as a Medical Device (SaMD)

The FDA's SaMD framework addresses AI-enabled medical devices through its Total Product Lifecycle (TPLC) approach, including pre-market review for diagnostic algorithms. However, post-market surveillance depends on manufacturer-reported adverse events, not independent behavioral monitoring. Stanford HAI researchers have warned that no well-articulated testing process exists for healthcare AI, and that AI tools are tested only by their developers. The 2024 analysis of 51 court cases involving software-related patient injuries in drug management, clinical decision support, and surgical robotics demonstrates that pre-market approval processes have not prevented harmful behavior in deployment.

NHTSA Automated Driving Systems Framework

NHTSA's framework for automated driving systems relies on voluntary safety self-assessments (VSSA), manufacturer self-reporting under the Standing General Order, and post-hoc recall authority. It does not include pre-deployment behavioral verification or continuous conformance monitoring. The Waymo school bus incidents demonstrate this gap: the system passed all development-phase testing, yet in deployment, it systematically violated traffic safety laws around school buses across multiple cities. NHTSA's enforcement mechanism—the recall—is reactive by definition, activated only after behavioral failures have already occurred in the field.

ISO 26262 and ISO/PAS 21448 (SOTIF)

ISO 26262 addresses functional safety for road vehicles, focusing on systematic failures in electrical and electronic systems. ISO/PAS 21448 (Safety of the Intended Functionality) addresses performance limitations and reasonably foreseeable misuse. Both standards are applied during the development process to identify and mitigate hazards. Neither provides a mechanism for continuous verification that the deployed system operates within the boundaries established during development. The gap between development-phase safety analysis and deployment-phase behavioral conformance is precisely where autonomous system failures occur.

3. The Comprehensive Comparison

The following table provides a systematic comparison of what each framework verifies, mapped against the specific requirements for autonomous system behavioral assurance:

Verification Domain	ISO 42001	NIST AI RMF	EU AI Act	SOC 2	ODDC
Organizational governance	✓ Full	✓ Full	✓ Full	✓ Full	— Not in scope
Risk assessment process	✓ Full	✓ Full	✓ Required	✓ Partial	— Not in scope
Technical documentation	✓ Required	✓ Guidance	✓ Required	✓ Required	✓ ODD specification
Operational boundary specification	— Not required	— Guidance only	— Implicit	— Not addressed	✓ Machine-readable ODD
Runtime boundary enforcement	— Not addressed	— Not addressed	— Not specified	— Not addressed	✓ ENVELO Interlock
Continuous behavioral verification	— Not addressed	— Not addressed	— Intent without mechanism	— Not addressed	✓ CAT-72 testing
Tamper-evident conformance records	— Not addressed	— Not addressed	— Not specified	— Audit logs only	✓ Cryptographic attestation
Independent third-party verification	✓ Accredited CABs	— Voluntary	✓ Required for high-risk	✓ CPA firms	✓ Sentinel Authority
Cross-manufacturer comparability	✓ Standardized MSS	— Not designed for	— Standards not ready	✓ Trust criteria	✓ Standardized status

The table reveals a clear pattern: existing frameworks thoroughly address organizational governance (the upper rows) while leaving behavioral verification (the lower rows) unaddressed. This is not a criticism—these frameworks were designed for organizational assurance. The gap exists because no framework was designed for autonomous system behavioral assurance. ODDC fills exactly this gap.

4. Real-World Consequences of the Gap

The distinction between process attestation and behavioral attestation is not academic. It has produced documented failures across multiple safety-critical domains.

4.1 Healthcare: Governance Without Behavioral Guardrails

ECRI, the independent healthcare safety organization, ranked AI systems deployed without proper oversight as the number one health technology hazard for 2025, and insufficient AI governance as the number two patient safety threat. These rankings were issued despite the widespread adoption of governance frameworks by healthcare organizations. The issue is not that hospitals lack AI policies—it is that those policies do not translate into enforceable constraints on AI system behavior.

The Department of Justice's subpoenas of pharmaceutical and digital health companies over AI deployed in electronic medical record systems targeted situations where governance documentation existed but the systems themselves produced harmful outputs. Stanford HAI researchers found that no well-articulated testing process exists for healthcare AI and that these tools are tested only by their developers—the very definition of a process-only assurance model.

The 2024 analysis of 51 court cases involving software-related patient injuries spans drug management systems that recommended dangerous dosages, clinical decision support tools that missed critical diagnoses, and surgical robotics that performed outside intended parameters. In each category, the deploying organization could point to governance policies, risk assessments, and vendor documentation. None of these process artifacts prevented the behavioral failure.

4.2 Autonomous Vehicles: Testing That Missed Deployment Behavior

Waymo's autonomous vehicles are among the most extensively tested in the industry, with over 100 million miles of autonomous driving and a published safety framework that references both internal testing and independent research partnerships. The Swiss Re study documented an 88% reduction in property damage claims compared to human drivers. By any process-attestation standard, Waymo's governance of its autonomous driving system would receive high marks.

Yet in deployment, Waymo's vehicles systematically violated school bus traffic safety laws across multiple cities. The Austin Independent School District documented 20 separate incidents during the 2025–2026 school year, including one where a student was still in the road when a Waymo vehicle passed the stopped bus. Five of these incidents occurred after Waymo had deployed software updates it believed would fix the problem. NHTSA's recall ultimately affected 3,067 vehicles—all running the same software that had passed development-phase testing.

This pattern—passing all process-based evaluations during development, then exhibiting systematic behavioral failures in deployment—is precisely what behavioral attestation is designed to detect. A system that has been verified to operate within its declared ODD through continuous conformance testing would have flagged the school bus recognition failure before it produced 20 incidents involving children.

4.3 Industrial Robotics: Documentation That Didn’t Prevent Injury

In September 2025, a lawsuit was filed against Tesla and Fanuc after a robotic arm struck a worker with approximately 8,000 pounds of force at Tesla’s Giga Texas facility, leaving him unconscious and requiring hospitalization. The $51 million suit alleges that the robot operated outside its programmed parameters. OSHA data spanning 2015–2022 documents 77 robot-related workplace accidents resulting in 93 injuries including amputations, fractures, and crushing injuries. NIOSH recorded 61 robot-related fatalities between 1992 and 2015.

Industrial robotics installations operate under extensive process documentation: risk assessments per ISO 12100, safety-rated control systems per ISO 13849, integration requirements per ANSI/RIA R15.06. The Tesla-Fanuc case illustrates that complete process documentation does not prevent a robot from operating outside its programmed parameters. Runtime behavioral enforcement—verification that the physical system actually respects the boundaries defined in the safety documentation—is the missing layer.

5. How ODDC Completes the Stack

5.1 The Complementary Architecture

ODDC is not designed to replace existing governance frameworks. It is designed to provide the behavioral verification layer that makes them complete. The architecture is explicitly complementary:

ISO 42001 + ODDC: ISO 42001 certifies that the organization has a management system governing AI development and deployment. ODDC certifies that the deployed system’s behavior conforms to the specifications documented in that management system. Together, they provide end-to-end assurance from organizational governance to runtime behavior.

NIST AI RMF + ODDC: The AI RMF’s Govern function establishes policies, Map identifies risks, Measure evaluates trustworthiness dimensions, and Manage implements risk responses. ODDC provides the verification layer that confirms the risk responses are actually effective at constraining system behavior. This directly addresses NIST AI 100-5e2025’s identified need for verification and validation tools and conformity assessment procedures.

EU AI Act + ODDC: The AI Act requires providers to demonstrate that high-risk systems meet accuracy, robustness, and cybersecurity requirements. ODDC’s formal ODD specification, ENVELO Interlock enforcement, and CAT-72 testing provide a concrete mechanism for generating the conformity evidence the Act demands—evidence that the missing harmonized standards have not yet defined how to produce.

SOC 2 + ODDC: SOC 2 verifies infrastructure controls around the AI system. ODDC verifies behavioral constraints within the AI system. The combination provides assurance that both the environment and the system itself are operating as intended.

5.2 The Three Verification Layers

ODDC’s behavioral attestation operates through three interlocking verification layers:

Layer 1 — Formal ODD Specification: The Operational Design Domain is specified in machine-readable format, defining the exact environmental conditions, performance parameters, and behavioral boundaries within which the system is designed to operate. This specification is not a marketing document—it is a formal contract between the system developer and the certification authority. Any deviation from this specification constitutes a conformance violation.

Layer 2 — ENVELO Interlock (Non-Bypassable Enforcement): The ENVELO Interlock is a hardware-software enforcement mechanism that prevents the autonomous system from operating outside its declared ODD. Critically, the Interlock cannot be disabled or overridden by the system it governs. This is the behavioral enforcement layer—not a policy that says the system should stay within boundaries, but a verified mechanism that prevents it from leaving them.

Layer 3 — CAT-72 Continuous Conformance Testing: The 72-hour Continuous Autonomous Testing protocol subjects the system to sustained operation under varied conditions to verify that behavioral conformance is maintained over time. Test results are cryptographically signed, independently stored, and tamper-evident. This provides the continuous verification dimension that no existing framework addresses.

6. The Standards Development Opportunity

NIST AI 100-5e2025 categorizes conformity assessment with AI standards as a Tier 2 priority—needed but requiring more scientific work or maturity before standardization. The plan explicitly calls for creating verification and validation tools and conformity assessment procedures as essential prerequisites for effective AI standards implementation. It further notes that standards are most useful if they arrive before technologies have moved on and warns that standards arriving too late risk failing to gain market acceptance.

The EU AI Act’s harmonized standards gap creates an immediate market opportunity: organizations need conformity evidence before the standards defining how to produce it are finalized. CEN/CENELEC’s missed deadline and the Commission’s proposed timeline extension demonstrate that institutional standards development is proceeding more slowly than the regulatory calendar demands.

ODDC represents a pragmatic bridge: a deployable conformance verification methodology that can be applied now, refined as standards mature, and ultimately incorporated into harmonized standards frameworks. For standards bodies, ODDC provides a working reference implementation of behavioral conformity assessment—exactly the kind of practical foundation that accelerates standardization.

For organizations facing imminent EU AI Act compliance deadlines, ODDC provides auditable behavioral evidence while the formal standards ecosystem catches up. For NIST, ODDC addresses the Tier 2 conformity assessment gap by providing tools and procedures that can be evaluated, refined, and eventually standardized through established processes.

7. Conclusion

The current AI governance ecosystem has achieved something remarkable: a global consensus that AI systems require organizational oversight. ISO 42001, the NIST AI RMF, the EU AI Act, and SOC 2 have collectively established that organizations must govern AI responsibly. This achievement should not be diminished.

But governance without behavioral verification is incomplete. A hospital with perfect AI governance documentation can still deploy a clinical decision support system that harms patients. A fleet operator with comprehensive safety processes can still deploy autonomous vehicles that fail to stop for school buses. An industrial facility with complete risk assessment documentation can still operate robots that exceed their safety parameters.

The gap between process attestation and behavioral attestation is the defining challenge of autonomous system safety. It is the gap that ECRI identifies when it warns about AI without proper oversight. It is the gap that NIST acknowledges when it identifies conformity assessment as a Tier 2 priority requiring more work. It is the gap that the EU AI Act creates when it mandates conformity evidence that harmonized standards have not yet defined how to produce.

NIST AI RMF tells you how to govern. ISO 42001 certifies that governance is in place. The EU AI Act requires conformity evidence. ODDC certifies that the system actually behaves as governance requires. This is not a competing standard—it is the missing layer that makes every other framework enforceable at the system level.

REFERENCES

[1] ISO/IEC 42001:2023. "Artificial Intelligence — Management System." International Organization for Standardization, December 2023.

[2] NIST. AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1. January 2023, with expanded companion playbooks through 2025.

[3] NIST. "A Plan for Global Engagement on AI Standards." NIST AI 100-5e2025. April 2025. Conformity assessment identified as Tier 2 priority.

[4] European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act). Entered force August 1, 2024. High-risk obligations applicable August 2, 2026.

[5] European Commission. Digital Omnibus Proposal, November 19, 2025. Proposed timeline extension linking high-risk rules to standards availability.

[6] CEN/CENELEC. prEN 18286: AI Quality Management System. First harmonized AI standard for EU AI Act. Entered public enquiry October 30, 2025.

[7] European Commission. General-Purpose AI Code of Practice. Released July 10, 2025.

[8] CSA. 2025 Compliance Benchmark Report: 76% of organizations plan to pursue ISO 42001 or similar frameworks.

[9] UNIDO. "Overview of ISO/IEC 42001 and AI System Conformity Assessment." Presentation on gap between digital service and tangible product conformity ecosystems. 2025.

[10] ISACA. "ISO 42001: Balancing AI Speed & Safety." October 2025. Analysis of framework as fast track to EU AI Act readiness.

[11] ECRI. "Top 10 Health Technology Hazards for 2025." AI without proper oversight ranked #1. Patient Safety Report: insufficient AI governance ranked #2.

[12] Stanford HAI. "Assessment of Trustworthy AI in the Context of Healthcare." Finding: no well-articulated testing process for healthcare AI; tools tested only by developers.

[13] U.S. Department of Justice. Subpoenas issued to pharmaceutical and digital health companies over AI in EMR systems. 2024.

[14] Bates, D. et al. "Software-Related Patient Injuries: Analysis of 51 Court Cases." Drug management, clinical decision support, surgical robotics. 2024.

[15] Waymo Recall of 3,067 Robotaxis. NHTSA investigation into school bus safety violations in Austin and Atlanta. December 2025.

[16] Austin Independent School District. Letter to Waymo and NHTSA documenting 20 illegal school bus passes during 2025–2026 school year.

[17] Tesla-Fanuc $51M Lawsuit. Robotic arm struck worker with ~8,000 lbs force at Giga Texas. September 2025.

[18] OSHA. Robot-Related Workplace Accidents: 77 incidents resulting in 93 injuries (2015–2022).

[19] NIOSH. Robot-Related Fatalities in the United States: 61 deaths (1992–2015).

[20] Swiss Re and Waymo. "Comparative Safety Performance of Autonomous- and Human Drivers." 25.3 million miles, 88% reduction in property damage claims. December 2024.

[21] NHTSA. Standing General Order on ADS Incident Reporting. Tesla ADAS vehicles: 53.9% of reported incidents.

[22] Frontier Model Forum. Comments on NIST TEVV Standard Outline. September 2025.

[23] Sentinel Authority. ODDC Overview v3.0, ENVELO Requirements v3.0, CAT-72 Procedure v3.0. Published at sentinelauthority.org.

This document references publicly available government reports, court filings, regulatory publications, and peer-reviewed research. This document does not constitute legal advice.