ThinkCapital LLC

Government IT / AI Governance Initiative

Implementation Fidelity:

Why AI RMF Adoption Metrics Are Measuring the Wrong Thing

ThinkCapital GIAG Research Series | March 2026

The compliance score problem

Federal agencies are reporting AI governance progress in terms that do not address the most pressing issue. The issue is not whether an agency has adopted the NIST AI Risk Management Framework. It is whether the governance function the framework is supposed to create is actually operating. Those are different questions, and current measurement practice conflates them.

OMB Memoranda M-25-21 and M-25-22 accelerate federal AI adoption while requiring governance controls as a condition of that acceleration.[1,2] The policy intent is sound. The measurement challenge it creates is not yet resolved: how does an oversight body, an agency head, or a congressional appropriator know whether the governance controls are working, as opposed to merely documented?

Current AI RMF adoption metrics count governance documentation activity. How many risk categories were addressed. How many risk tier assignments were completed. How many designated AI points of contact were named. These metrics are internally consistent and auditable. They also measure the governance equivalent of lines of code: technically precise, functionally uninformative about what the governance system delivers.

The software measurement community confronted an identical structural problem in the 1980s, when productivity metrics based on lines of code produced figures that described implementation volume while telling managers nothing reliable about functional delivery. The resolution came from identifying the wrong unit of measurement and replacing it with one anchored to what the system actually produced for its users.[3] The same move is required in AI governance, and the conceptual tools to make it are available.

Defining implementation fidelity

Implementation fidelity is the degree to which a governance framework changes actual decision behavior. It is distinct from three other concepts that current measurement practice treats as proxies for it.

Three concepts that are mistaken for implementation fidelity

Adoption rate is whether the framework has been formally accepted and assigned. An agency can adopt a framework on day one and operate it at zero fidelity indefinitely.

Compliance score is whether required documentation exists and required process steps were completed. A process that produces completed checklists without influencing the decisions those checklists are meant to govern is compliant and non-functional.

Maturity rating is where an agency self-assesses on a capability scale. Maturity models measure process definition, not process performance. A well-defined governance process that is systematically bypassed in practice scores higher than a less-defined process that is consistently applied.

A high-fidelity governance implementation is one where the framework's presence demonstrably changes what decisions are made and how. A low-fidelity implementation is one where the framework's documentation exists, but decision behavior is indistinguishable from what it would have been without the framework. The NIST AI RMF is explicit on this distinction: it separates the governance function (the organizational capability to manage AI risk) from governance documentation (the evidence that the function exists).[4] Current federal reporting metrics have not made that translation operational.

Three agencies, identical compliance scores

Consider three federal agencies, each of which reports full AI RMF adoption, documented risk tier assignments across their AI portfolio, and a designated AI governance official.

Agency A has completed all required documentation. Risk tier assignments were produced by the same team that developed the systems being assessed, reviewed for form rather than substance, and approved without material challenge. No AI deployment in the past 18 months has been modified because of the governance process. The compliance score is exemplary.

Agency B has completed all required documentation. Risk tier assignments were produced by an independent review function that challenged two high-risk system classifications, resulting in one deployment delay and one significant redesign of a decision support interface. One post-deployment monitoring finding triggered a documented constraint on the system's use in a specific case type. The compliance score is identical to Agency A's.

Agency C has incomplete documentation. Risk tier assignments are current for sixty percent of the portfolio. The agency's review function, though under-resourced, has produced five

governance challenges in eighteen months, three of which modified deployment decisions. The compliance score is substantially lower than Agencies A or B.

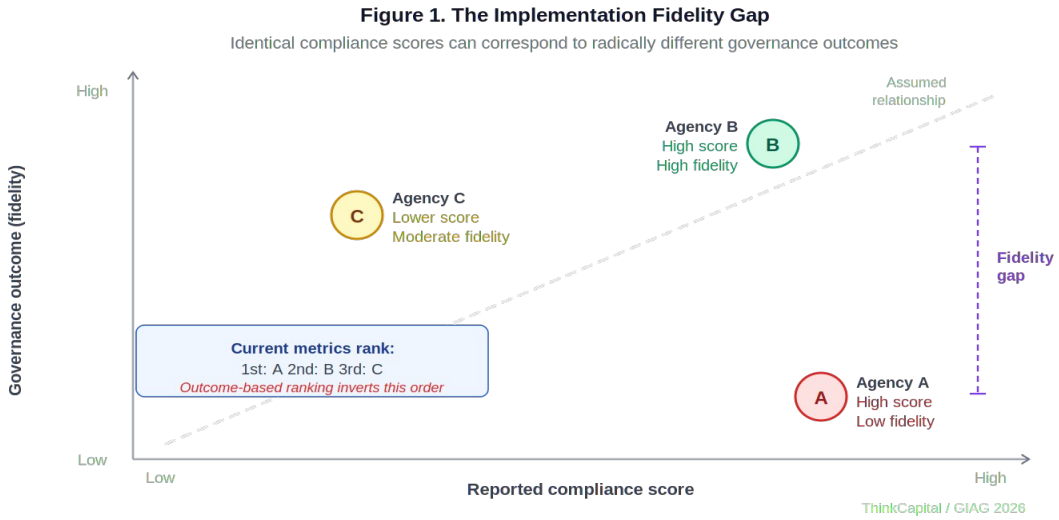

Figure 1. The Implementation Fidelity Gap Agencies A, B, and C occupy very different positions on a governance outcome axis despite near-identical compliance scores. Current metrics rank A first, B second, and C third. On any outcome-based assessment, the order inverts (see Figure 3). The fidelity gap is the distance between the assumed and actual relationship between score and outcome.

Current measurement frameworks rank Agency A highest, Agency B second, and Agency C third. Any outcome-based assessment reverses that order. This is not an edge case or a theoretical anomaly. It is the predictable result of measuring activity at the documentation layer while the governance function operates, or fails to operate, at the decision layer.

What fidelity-based measurement would actually track

A measurement framework designed to assess implementation fidelity rather than adoption activity tracks outcomes at the decision layer, not the documentation layer. The operational indicators are not complex to define. They require governance processes that produce traceable decision records rather than completed checklists, and oversight functions that have access to those records and the authority to act on them.

Four core metrics constitute a minimum viable fidelity measurement set.

- **Review efficacy rate:** The proportion of AI governance reviews that produce a documented outcome other than unconditional approval. A governance process in which every review results in approval without modification is either reviewing systems that genuinely require no modification or not reviewing them substantively. Neither high-risk system portfolios nor mixed-maturity development organizations produce portfolios with zero substantive governance findings.
- **Classification integrity rate:** The proportion of initial risk classifications that survive independent challenge without change. High initial agreement between system developers and governance reviewers, sustained across all systems without challenge, is a signal of process capture rather than sound risk assessment. Independent challenge is the mechanism by which classifications carry meaning.
- **Operational constraint compliance:** The proportion of systems operating under risk-tier-appropriate constraints, verified through operational monitoring rather than documentation review. A system classified as high-risk and approved with constraints is not governed if those constraints are not enforced in production and verified through an active monitoring process.
- **Monitoring response rate:** When post-deployment monitoring produces findings that exceed defined risk thresholds, the proportion of those findings that trigger documented responses within defined timeframes. Monitoring that generates findings without documented responses is instrumentation, not governance.

Figure 2. From Adoption Metrics to Fidelity Metrics

What current measurement counts vs. what implementation fidelity requires

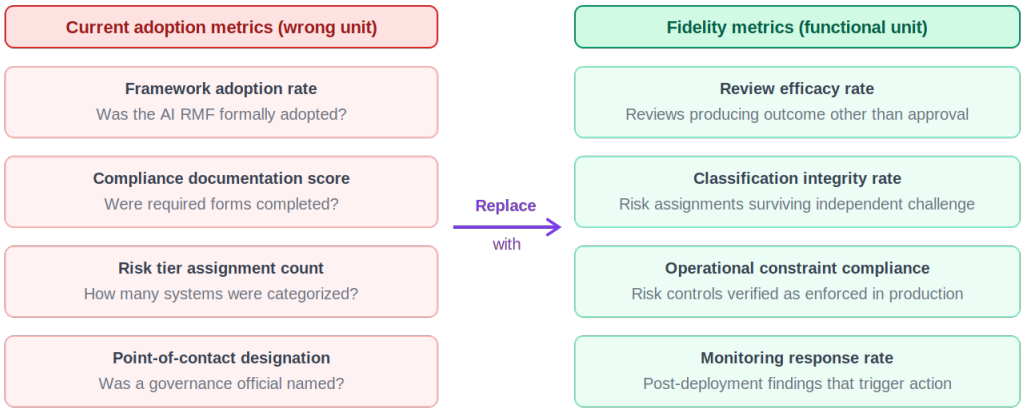

Figure 2. From Adoption Metrics to Fidelity Metrics Current adoption metrics (left) measure activity at the documentation layer. Fidelity metrics (right) measure outcomes at the decision layer. The replacement is not a technical challenge; it is an organizational design choice about what governance processes are required to produce and record.

These four metrics share a structural requirement: the governance process must produce traceable records of decisions and their outcomes, not just evidence that a process step was completed. An agency whose governance process generates documentation of reviews conducted, challenges made, classifications revised, constraints assigned, and monitoring

findings actioned has the raw material for fidelity measurement. An agency whose process generates only completion confirmations does not, regardless of its compliance score.

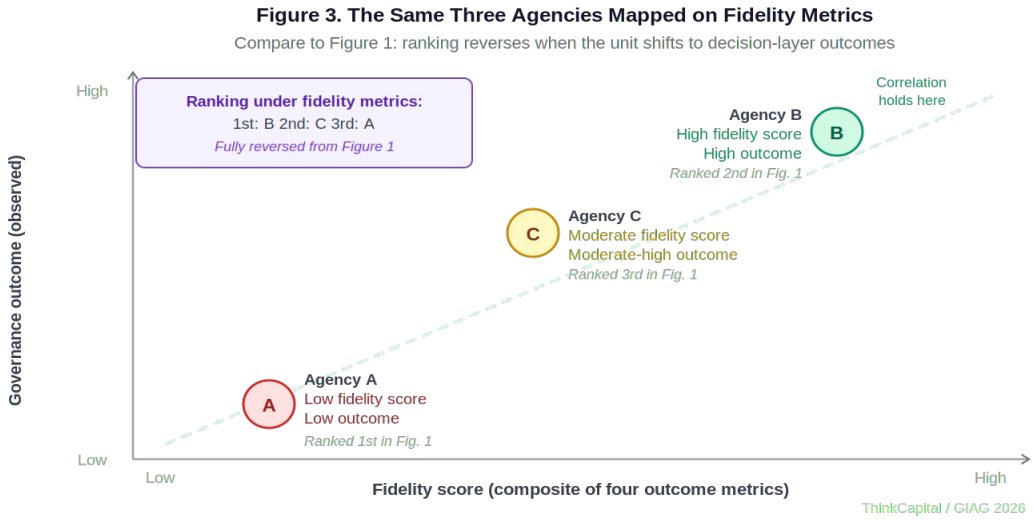

Figure 3. The Same Three Agencies Mapped on Fidelity Metrics Compare to Figure 1. When the unit of measurement shifts from documentation activity to decision-layer outcomes, the ranking of Agencies A, B, and C reverses completely. Agency A, the highest compliance scorer, maps to the lowest fidelity position. Agency B holds its position. Agency C, the lowest compliance scorer, moves ahead of A. The fidelity metrics produce a governance signal that compliance scores cannot.

The policy context: what current OMB guidance establishes and what it leaves open

OMB M-25-21 directs agencies to remove unnecessary barriers to AI adoption while maintaining appropriate risk controls.[1] M-25-22 provides acquisition guidance for AI procurement, requiring agencies to evaluate AI systems against defined governance criteria before and after deployment.[2] Both memoranda accelerate the operational deployment of AI systems under governance structures that the NIST AI RMF describes.[4]

What neither memorandum specifies is how oversight bodies should assess whether those governance structures are functioning. The NIST AI RMF distinguishes between the governance function and governance documentation throughout its core guidance. The framework does not treat documentation as a substitute for function. Current federal reporting mechanisms have not operationalized that distinction.

The measurement gap this creates is not administrative. It is a structural accountability gap. Agencies can report full compliance with AI governance requirements while operating governance functions of near-zero fidelity. Without fidelity-based measurement, no oversight body, whether OMB, agency inspectors general, or appropriations committees, has the instruments to tell the difference. The policy infrastructure for AI governance is further developed than the measurement infrastructure for assessing whether that governance is working.

GIAG Stream One: structured inquiry into the fidelity gap

The Government IT and AI Governance Initiative is conducting structured research designed to produce the first empirical data on this gap across government agencies. Stream One of the GIAG research examines NIST AI RMF implementation fidelity, asking a specific question: for agencies that report AI RMF adoption, what is the measurable relationship between reported compliance scores and governance outcomes at the decision layer?

The research protocol collects documented evidence of governance process activity, including review records, classification challenge logs, deployment modification records, and monitoring response documentation, and maps that evidence against reported compliance metrics. The measurement framework draws on the functional sizing methodology described in the companion technical methods paper,[5] extending it to treat governance process outputs as the functional unit of measurement rather than governance documentation activity.

The research is designed to produce three outputs. First, an empirical calibration of the relationship between compliance scores and governance outcomes across a representative sample of government agencies. Second, a validated fidelity measurement protocol that agencies and oversight functions can apply without specialized measurement expertise. Third, a set of structural findings about which governance design features are associated with higher fidelity, independent of framework compliance level.

Research participation

If you manage AI governance programs in a federal, state, or local government context and have access to governance process records, including review outcomes, classification challenge histories, deployment modification records, or monitoring findings, the GIAG research team is interested in your experience. Positive cases and negative cases are both analytically valuable. Participation is confidential at the agency level; findings are reported in aggregate.

Research participation details and enrollment: thinkcapital.org/research.html

What the field needs next

The governance infrastructure built around the NIST AI RMF over the past three years represents a substantial investment in organizational capacity. The question now is whether that investment is producing what it was designed to produce, and no current measurement system can answer it reliably.

Answering it requires a shift in the unit of measurement from documentation activity to decision-layer outcomes. That shift is technically straightforward and organizationally demanding. It requires governance processes that produce traceable decision records, oversight functions that have access to those records, and reporting frameworks that ask about outcomes rather than completion rates.

The measurement methodology for making that shift at scale, extending functional sizing principles to AI governance process outputs, is developed in the companion technical methods paper and is the subject of ongoing empirical research through GIAG. What the field needs is not more compliance scores. It is calibrated evidence about whether the governance machinery that has been built is producing the outcomes it was designed to produce.

References

[1] Office of Management and Budget. (2025). *Memorandum M-25-21: Accelerating Federal Use of AI through Streamlined Governance.* Executive Office of the President, Washington, DC. March 2025.

[2] Office of Management and Budget. (2025). *Memorandum M-25-22: Driving Efficient Acquisition of Artificial Intelligence in Government.* Executive Office of the President, Washington, DC. March 2025.

[3] Jones, C. (2008). *Applied Software Measurement: Global Analysis of Productivity and Quality,* 3rd ed. McGraw-Hill. See also Albrecht, A.J. (1979). “Measuring Application Development Productivity.” *Proceedings of the IBM Applications Development Symposium, SHARE/GUIDE,* Monterey, CA.

[4] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* NIST AI 100-1. U.S. Department of Commerce, Gaithersburg, MD.

[5] Bragen, M. (2026). *Functional Sizing as a Foundation for AI Governance Measurement: Applying Function Point Analysis and COSMIC to AI System Scope and Complexity.* Technical Methods Paper No. 1, ThinkCapital GIAG Research Series. ThinkCapital LLC, Belmont, CA.