AI Governance – The Challenge Facing Federal Leaders

Robert Dowling

Table of Contents

Table of Figures

1.0 Independence as a First Principle in AI Governance

One of the most consistently observable properties of deployed AI systems is change over time. Even when no code is modified, no parameters are retrained, and no explicit release is performed, system behavior shifts. Inputs evolve, user interaction patterns change, prompt structures drift, and the statistical properties of real-world data diverge from the conditions under which the system was initially evaluated and accepted for deployment.

In operational terms, this phenomenon is most commonly surfaced as model **drift**.

Drift is not inherently a defect. In many cases, it is the differentiating feature. Systems operating in live environments must adapt to changing conditions to remain useful and effective within evolving workflows. The problem is not that drift exists; the problem is that drift often goes undetected, unclassified, and ungoverned.. When drift is detected, it is frequently ascribed as a technical or performance issue rather than a behavioral shift with direct governance implications. In other cases, drift appears only in fringe outputs and the default response becomes "retrain the model"—driving cost, increasing operational friction, and obscuring the underlying cause.

This matters because drift is rarely isolated. It is often the first visible indicator of deeper behavioral change. A system that begins to drift statistically may also begin to exhibit subtle shifts in confidence, boundary adherence, refusal behavior, or susceptibility to user pressure. These changes may not immediately manifest as errors or failures, but they alter how the system behaves relative to its authorized purpose.

In most Federal AI deployments today, the responsibility for identifying, interpreting, and responding to these signals rests with the same organization that built the system, or with an oversight body not embedded in the technical and operational realities of the large language model (LLM) deployment. This creates a structural problem that has nothing to do with competence or intent. Developers are incentivized to optimize performance, maintain uptime, and demonstrate mission value. Governance, by contrast, requires the willingness to surface degradation, uncertainty, and risk—to slow down delivery, pause operations, and raise uncomfortable questions about system readiness. Oversight entities often operate across broad portfolios and lack the time, access, or telemetry required to manage operational AI risk in detail. They are asked to govern what they cannot observe.

This tension is not unique to AI. History is littered with technologies that outpaced their oversight mechanisms. Federal acquisition has long recognized that independent oversight is required when systems become complex, adaptive, and mission-critical. Developers do not certify their own cloud environments. They do not audit their own financial controls. They do not independently validate their own cybersecurity posture. In each of these domains**, independence emerged not as a policy preference, but as a structural imperative**—born from repeated failure, once systems exceeded a threshold of complexity and risk.

AI systems have crossed that threshold. We are now in a period where oversight independence is optional. The coupled opportunity and risk profile of AI suggests this should not last.

What makes AI different is not simply that models are opaque or probabilistic. It is that behavior evolves continuously under real-world use, often in ways that are not captured by traditional testing or acceptance processes. Drift is simply the most measurable expression of this reality. Other manifestations—including hallucination, bias amplification, prompt-induced behavior change, passive agreement, or unjustified refusal—are harder to quantify but arise from the same

underlying condition: the system is diverging from its authorized bounds, and no one is watching.

When developers are tasked with governing these behaviors, the result is not malfeasance; it is systemic blind spots. Subtle changes are normalized as expected variance. Edge cases are rationalized as acceptable tradeoffs. Signals that do not map cleanly to performance metrics are deprioritized. Over time, the system's behavior diverges from its original authorization envelope without any clear inflection point at which governance intervenes.

This is why **independence must be treated as a first principal** in AI governance, not an optional enhancement deferred to the future. The more that is built around a diverging LLM, the larger and harder-to-unwind the risk surface area becomes. Every day a drifting system operates undetected is another layer of organizational dependency, another integration point, another sunk cost. Continuous assurance cannot be an extension of development operations, and it cannot be reduced to periodic evaluation. It must exist as a function whose success is measured not by deployment velocity or model accuracy, but by its ability to observe, classify, and escalate behavioral change—starting with drift, but never ending there.

2.0 Why Federal AI Governance Has Not Yet Reached the System Level

At the policy level, Federal AI governance appears increasingly well defined. Executive Orders, OMB memoranda, and risk management frameworks articulate expectations around accountability, transparency, and responsible use. Agencies are instructed to inventory AI systems, assess risk, and establish governance structures. From a distance, it would be reasonable to assume that operational assurance mechanisms already exist to support these mandates.

At the system level, however, that assumption does not yet hold.

In practice, Federal AI governance remains largely **declarative rather than operational**. It articulates what agencies should care about, but not how those concerns are continuously measured, enforced, or independently validated once a system is deployed. Governance expectations are commonly expressed as principles rather than as measurable, enforceable requirements tied to observable behavior.

This gap is not surprising. In other technology domains, governance matured only after repeatable implementation patterns emerged. FedRAMP translated cloud security policy into standardized authorization processes, continuous monitoring requirements, and independent assessment roles. Cybersecurity governance evolved through continuous diagnostics, separation of duties, and third-party validation. Financial systems rely on independent audit as a matter of course. In each case, policy became operational only when mechanisms existed to measure compliance continuously and assign accountability clearly.

AI has not yet completed this transition.

One reason is timing. AI systems moved from experimentation to operational deployment faster than governance mechanisms could adapt. Another is ambiguity. Many AI risks are behavioral rather than deterministic, making them difficult to encode as static controls or checklist items. Most significantly, existing governance models were designed for systems whose behavior remains relatively stable between reviews. AI systems—particularly those built on LLMs continuously adapt and evolve. They do not behave that way.

As a result, governance responsibilities are often fragmented. Developers are expected to self-monitor. Program offices rely on periodic reporting. Oversight boards focus on policy alignment

rather than operational telemetry. Each function acts in good faith, yet the structure itself ensures that no entity is positioned to continuously observe how a system behaves under real-world use.

Until agencies have concrete, implementable models for truly independent, continuous AI assurance, governance will remain aspirational—well articulated in guidance but inconsistently executed in practice.

3.0 From Drift to Behavioral Integrity

Drift is often the first signal that draws attention in a deployed AI system because it is measurable, comparative, and difficult to ignore. Statistical divergence between baseline and current behavior can be quantified, trended, and thresholded. For many teams, drift becomes the entry point into operational monitoring precisely because it aligns with familiar analytical techniques.

But drift is not the risk. It is an indicator that the system is changing.

Focusing on drift alone can create a false sense of sufficiency. A system may remain statistically stable while its behavior degrades in more subtle ways. Conversely, a system may exhibit measurable drift while remaining fully aligned with its intended purpose. Treating drift as the primary governance signal risks confusing change with harm, and stability with safety.

The most consequential AI failures rarely stem from distributional shift alone. They emerge from changes in how a system thinks, responds, and behaves under pressure—changes that may not register immediately in aggregate statistics.

Hallucination is the most visible example. It attracts attention because it represents an extreme and easily recognizable failure mode: confident outputs unsupported by evidence. But hallucination is not an isolated defect. It is a behavioral expression of deeper alignment pressures. LLMs are optimized to be responsive, helpful, and fluent. Under ambiguous or adversarial conditions, that optimization can manifest as speculative reasoning or fabricated detail.

From a governance perspective, hallucination matters less as a category than as a signal. It reveals how the system behaves when the boundary between “helpful” and “accurate” becomes unclear.

The same is true of other behavioral risks that receive less attention but are equally consequential. Systems may begin to defer without scrutiny. They may refuse without clear justification. They may show increased susceptibility to prompt injection, subtly shifting behavior in response to user input rather than stated intent. Bias may amplify as usage patterns shift, even when overall accuracy remains stable.

These behaviors share a common trait: they are not reliably captured by traditional performance metrics. They do not always trigger obvious failures. Instead, they accumulate gradually, altering the system’s decision posture relative to its authorized role.

Assurance must therefore focus on **behavioral integrity**, not individual failure modes. Drift, hallucination, bias, and injection susceptibility are not separate problems to be solved independently. They are different manifestations of the same underlying reality: an adaptive system operating under real-world conditions will inevitably change in ways that cannot be fully anticipated at design time.

These behavioral signals can be observed concurrently without access to model internals.

Figure 1: Representative behavioral integrity signals observed across multiple dimensions

No single indicator determines risk; interpretation depends on how signals move together over time.

4.0 Demonstrating the Feasibility of Independent, Continuous AI Assurance

This whitepaper is not intended solely to describe the challenges associated with governing deployed AI systems, nor to argue abstractly that a solution should exist. It was written to document a third and more consequential outcome: that independent, continuous AI assurance can be implemented in practice, under realistic Federal constraints, and deployed successfully to a live, learning AI system.

The work described here began with a hypothesis—that it should be possible to observe, measure, and govern AI system behavior over time without privileged access to model internals, without vendor cooperation, and without interfering with production inference. In other words, assurance could operate independently, at the system boundary, independent of development incentives, while still producing sufficient evidence for governance decisions.

To test this hypothesis, we designed an assurance framework explicitly oriented around independence, temporal analysis, and behavioral observation. The framework was not conceived as a policy construct or a reference architecture, but as an operational and adaptable solution: one capable of executing repeatedly, establishing authorized baselines, measuring behavioral change across successive runs, and producing interpretable signals suitable for governance oversight across a wide range of AI systems.

That framework was then instantiated as a working prototype and deployed against a functional LLM operating under learning conditions. The model was not static, and the environment was not artificially constrained to produce favorable results. The assurance system executed alongside normal operation, observing inputs, outputs, decisions, confidence signals, and control-plane actions without modifying model behavior or influencing outcomes.

Across repeated executions, the system successfully identified and tracked behavioral change over time, including—but not limited to—distributional drift, shifts in confidence coherence, changes in decision-path composition, and variations in response behavior under stress. These signals were measured longitudinally, contextualized against an authorized baseline, and presented as governance-relevant evidence rather than as binary pass / fail determinations. This is because many AI systems do not operate with ground truth due to resource constraints and the redundancy required in evaluating response accuracy.

The significance of this result is not tied to the specific use case. Email classification served as a practical and instrumentable environment, not as a limitation on applicability. The purpose of the prototype was to demonstrate that a governance approach—independent, continuous assurance grounded in behavioral observation—can be operationalized in a real AI system, under Federal Deployment constraints.

This paper therefore serves three purposes: to describe the governance gap that exists today, to explain the principles required to close it, and to document that those principles have been successfully implemented and validated in practice. The remaining sections focus on how this assurance model was designed, what architectural choices made it possible, and how similar approaches can be applied across other AI use cases.

5.0 Designing Assurance Without Access to the Black Box

Effective AI assurance does not require privileged access to model internals. In many Federal environments, such access is neither feasible nor desirable. In fact, the recent 2025-12-11 OMB Memorandum regarding AI acquisition states that agencies should avoid requirements that compel vendors to disclose sensitive technical data (e.g., specific model weights). Systems may rely on vendor-managed models, closed-source architectures, or externally hosted services. Governance that depends on introspection into the black box is therefore brittle by design.

Instead, assurance must operate at the **system boundary**.

By observing inputs, outputs, decisions, confidence signals, and control-plane actions, it is possible to infer meaningful behavioral patterns without interfering with model operation. This boundary-focused approach preserves independence while remaining compatible with a wide range of deployment architectures.

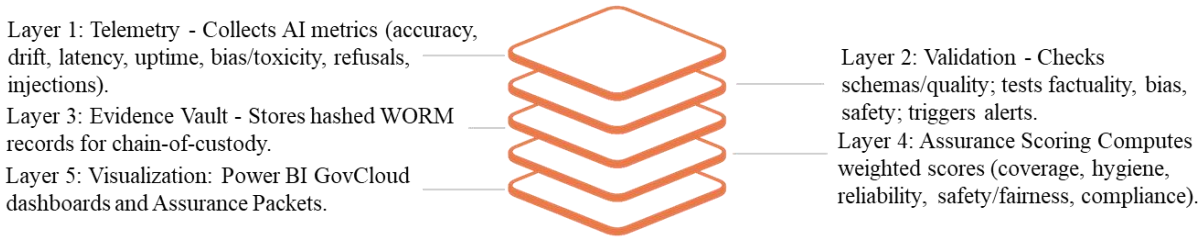

Figure 2: Separation between model execution, control-plane enforcement, and independent assurance

Critically, independence is not an accidental byproduct of this design choice—it is the objective. Assurance functions should not be embedded within development pipelines or dependent on developer discretion. They should observe, measure, and escalate without modifying production behavior. Read-only telemetry, immutable evidence, and clear separation of duties are not constraints; they are governance requirements.

This architectural posture allows assurance to remain viable even as models, vendors, and

deployment environments change.

6.0 Assurance as a Temporal Discipline

Traditional evaluation treats AI systems as static artifacts assessed at discrete points in time. Deployed AI systems are not static. Their behavior evolves continuously under real-world use.

Assurance must therefore be **temporal by design**.

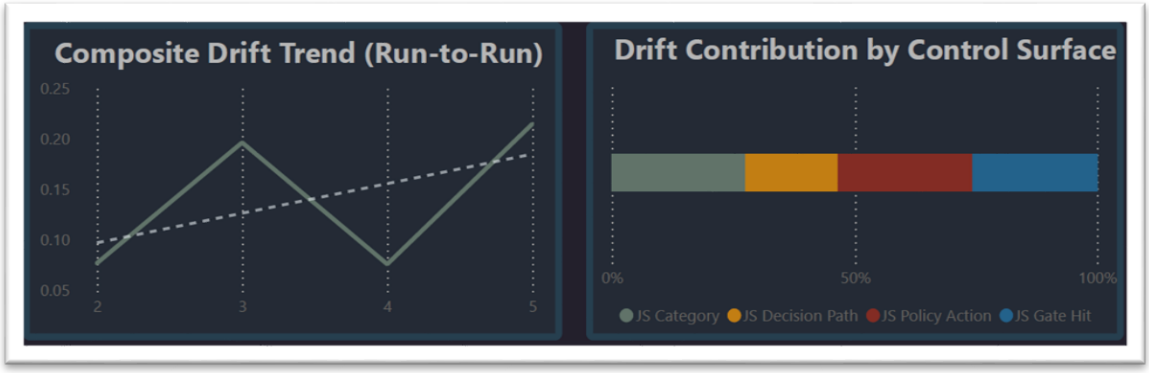

Figure 3: Representative drift measurement

The critical question is not whether a system performs acceptably at a given moment, but how its behavior changes over time relative to an authorized baseline. Temporal comparison—baseline to current, previous to current—reveals trends, inflection points, and step-changes that static evaluation cannot.

This longitudinal perspective enables governance responses that are proportional rather than reactive. Not all change warrants intervention. Some change reflects healthy adaptation. The role of assurance is to continually measure for change, then distinguish between adaptation and degradation.

7.0 Instrumenting Behavioral Signals Across Use Cases

No single metric can represent AI trustworthiness. The signals that matter most depend on the mission context.

In an email triage system, distributional drift, confidence coherence, and decision-path stability may dominate. In a socioeconomic forecasting model, bias, calibration, and fairness metrics may be primary. In a decision-support system exposed to untrusted inputs, adversarial pressure and boundary adherence may be critical.

The assurance framework must therefore adapt its telemetry without changing its core principles. Independence, temporal analysis, and behavioral interpretation remain constant; the signals emphasized vary by use case.

This adaptability is what allows assurance to scale across domains without becoming a bespoke, one-off exercise.

8.0 What This System Demonstrates—and What It Does Not

The prototype described in this paper demonstrates that continuous, independent AI assurance is technically feasible. Behavioral change can be observed without access to model internals. Drift can be contextualized rather than reflexively remediated. Multiple behavioral signals can be

tracked concurrently and interpreted over time.

It does not claim to eliminate risk, guarantee correctness, or replace human judgment. Assurance surfaces signals; governance decisions remain human responsibilities.

Drawing this boundary is essential. Credible governance is built to demonstrate capability, not aspirational claims.

9.0 Implications for Federal AI Acquisition and Oversight

As AI systems become embedded in mission-critical workflows, acquisition and oversight models must evolve. Governance requirements must move beyond principles and into measurable, operational expectations: independent monitoring, defined behavioral thresholds, evidence retention, and escalation authority.

Assurance should be specified as a function, not an afterthought—funded, staffed, and evaluated independently from development.

Without these mechanisms, agencies will continue to rely on trust and periodic reporting in environments that demand continuous visibility.

10.0 Conclusion: Governing Behavior, Not Models

AI governance is no longer a theoretical exercise. Deployed systems change continuously, often in subtle ways that escape traditional oversight. Drift is only the beginning.

The next phase of Federal AI governance will be defined not by new policy frameworks, but by the operationalization of independent, continuous assurance. Governance must shift from evaluating models to governing live behavior—clearly, independently, and continuously.

This is not a future requirement. It is a present necessity.