The Law of Evaluation: Beyond Benchmarks for AI Governance

JOURNAL OF LAW AND INNOVATION (2026, forthcoming)

Daniel E. Ho, Olivia H. Martin & Joyce Kasing Tagal[1]

Abstract

AI systems are growing more complex, more embedded, and more consequential all while the practices for evaluating these systems remain strikingly thin. Models now shape high-stakes decisions across private and public sectors, yet are often assessed through static benchmarks, internal demos, or short-term testing that reveal little about real-world performance, distributional effects, or downstream impact. This Article argues that effective AI governance must be built around legally grounded, ongoing evaluation rather than episodic or purely technical assessment. Drawing on public law frameworks governing performance measurement, evidence-based policymaking, and procurement, we show that existing law already supplies the mandate, authority, and framework for more rigorous, system-level evaluation of AI deployments. For evaluation to produce meaningful oversight, central institutional challenges of independence and resourcing must be overcome, coupled with necessary technical advances. Risk-tailored evaluation that integrates pilot testing, randomized trials, quality assurance, and performance measurement will also be central. We conclude by outlining institutional models through which AI researchers, funders, governments, and private actors can embed evaluation as a core feature of AI governance.

Introduction

When a fast-growing AI startup approached a prominent law school with an offer to become one of the first academic institutions to purchase a license for a new generative AI platform, both sides saw an easy win. The company would get a marquee institutional partner and a press release. The law school could signal that it was forward-looking and technologically sophisticated.

But buried in the licensing terms was a startling clause: The school—an elite institution with faculty experts in AI governance, a standing AI committee, and a mission to interrogate emerging technologies—was prohibited from conducting any "test or review" of the system. In

[1] Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, Professor of Computer Science (by courtesy), Senior Fellow at the Stanford Institute for Economic Policy Research (SIEPR) and the Stanford Institute for Human-Centered AI, and Faculty Director of the RegLab. Martin earned her J.D. at Stanford Law School and is a Ph.D. candidate in Economics at Stanford University and a Graduate Student Fellow at RegLab. Tagal is a Senior Data Scientist at Stanford RegLab. We are grateful to Christie Lawrence for incisive suggestions, Margot Kaminski for a lucid discussion of an earlier version of the paper, and Cary Coglianese, Chris Yoo, and participants at the *Journal of Law and Innovation* symposium on "Steering Complexity: Adaptive Governance for Emerging Technologies" for helpful comments.

the eagerness to be early adopters, the institution had nearly contracted away the most basic element of responsible governance: the ability to understand how the system behaved.[2]

The irony is difficult to overstate. If any institution were equipped to evaluate a cutting-edge AI system, it would be a law school filled with scholars of law and technology, contracts, AI governance, and empirical methods, not to mention students and clinics dedicated to studying algorithmic accountability. Yet even here, evaluation was nearly disallowed. The problem is not confined to academia. At one of the country's leading law firms, lawyers spent more than a year piloting a bespoke legal assistant built on generative AI. But public reporting suggests that this experimentation has proceeded without rigorous metrics tying the tool to concrete productivity gains or bottom-line performance, underscoring how even sophisticated private actors treat evaluation as optional, informal, and largely inward-facing.[3]

This vignette reflects a broader dynamic: AI adoption is accelerating far faster than evaluation practices, even in our most sophisticated institutions. AI systems are being procured and deployed across an ever-expanding array of domains. A 2024 McKinsey survey found that 78% of companies surveyed reported using AI in at least one business function, up from 55% in 2023.[4] Such a rate of adoption is unlikely to change so long as companies face external pressures, particularly from clients, to use generative AI as a perceived way to cut costs and timelines.[5] In 2024, federal agencies reported 1,700 AI use cases, more than double the number from 2023.[6] Of those, more than 10% were labeled "rights-" or "safety-impacting." A 2025 ICF survey of federal agencies found that 41% are running small-scale AI pilots, 16% are actively scaling AI efforts, but only 8% are seeing results from mature programs.[7] Yet the basic questions one might ask about any consequential system—Does it work as intended? For whom? At what cost and with what side effects?—often go unanswered.

Had the buyer in our opening vignette been a federal agency rather than a law school, the story might have looked quite different. A federal agency would be mandated by the U.S. Office of Management and Budget (OMB) – the department tasked with overseeing agency implementation of the President's policy agenda – to "measure, monitor, and evaluate" ongoing

[2] After last-minute negotiations, the usage restrictions were removed. That revision is itself suggestive: the school may have had more leverage than it expected, or the clause may have been boilerplate included by default in the vendor's form agreement. Either way, the episode underscores how much turns on diligence at the contracting stage, even for "standard" GenAI licenses.

[3] Isabel Gottlieb, *Paul Weiss Assessing Value of AI, But Not Yet on Bottom Line*, BLOOMBERG LAW (May 14, 2024), https://news.bloomberglaw.com/business-and-practice/paul-weiss-assessing-value-of-ai-but-not-yet-on-bottom-line.

[4] MCKINSEY & CO., THE STATE OF AI (MAR. 2025), https://www.mckinsey.com/~/media/mckinsey/business functions/quantumblack/our insights/the state of ai/2025/the-state-of-ai-how-organizations-are-rewiring-to-capture-value_final.pdf.

[5] *See e.g.*, David L. Brown, *Clients Are Demanding AI and Savings—Can Law Firms Deliver?*, Best Law Firms (Sept. 19, 2025), https://www.bestlawfirms.com/articles/clients-demand-ai-savings-can-law-firms-deliver/6910.

[6] Madison Alder, *Federal Government Discloses More Than 1,700 AI Use Cases*, FEDSCOOP (Dec. 18, 2024), https://fedscoop.com/federal-government-discloses-more-than-1700-ai-use-cases.

[7] *The AI Advantage: Moving from Exploration to Impact*, ICF (June 17, 2025), https://www.icf.com/insights/technology/data-ai-trends-federal-government-report-2025.

performance and risks in AI applications.[8] In procuring AI tools, agencies must ensure that contracts do not foreclose their compliance with OMB's obligations, including their ability to test and assess AI tools used in public-facing or high-stakes setting. [9] A federal contract that purported to forbid any "test or review" of a rights-impacting AI system would thus conflict with emerging procurement requirements.

AI governance debates have focused narrowly on benchmarking, audits, and other AI-specific assessment tools, but these approaches overlook a much broader tradition of legally required evaluation in public law. Despite strong incentives for under-evaluation in the private sector, public-sector AI policy—through recent executive orders and OMB guidance—has begun to move toward more meaningful oversight. Yet even these efforts remain partial. By situating AI evaluation within the wider landscape of federal performance measurement, evidence-based policymaking, and procurement quality assurance, we argue that effective AI governance requires reconnecting the emerging AI toolkit with longstanding evaluation obligations in administrative and procurement law. This broader perspective reveals two central design challenges—independence and resourcing—and points to institutional models capable, coupled with the right technical advances, of meeting them.

The rest of the Article proceeds as follows. Part I diagnoses why prevailing approaches to AI evaluation—benchmarking, internal pilots, red-teaming, impact assessments, and periodic audits—regularly fail to answer the questions that matter in real deployments, even as executive-branch guidance has begun to push agencies toward more systematic testing and monitoring. Part II recovers the broader landscape of public-law evaluation requirements—the Government Performance Results Act (GPRA), the Evidence Act, and procurement law—and assesses both their reach and their practical shortcomings when applied to AI systems. Part III then turns to institutional design, identifying concrete models that can secure independent, well-resourced AI evaluation and outlining needed contributions from technologists, lawyers, funders, and policymakers. Throughout the Article, we draw heavily on public law frameworks. However, our aim is to inform AI evaluation and governance across sectors—public and private.

I. Existing AI Evaluation Frameworks and Their Limits

A. Conventional AI Assessments

Within machine learning, evaluation has conventionally meant benchmarking, which measures a model's performance against standardized metrics and datasets. Benchmarking

[8] *See* Memorandum from Russel T. Vought, Dir., Office of Mgmt. & Budget, to the Heads of Exec. Dep'ts & Agencies, Accelerating Federal Use of AI through Innovation, Governance, and Public Trust, Memo M-25-21 (Apr. 2025), https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf (hereinafter OMB Memo M-25-21).

[9] *See* Memorandum from Russel T. Vought, Dir., Office of Mgmt. & Budget, to the Heads of Exec. Dep'ts & Agencies, Driving Efficient Acquisition of Artificial Intelligence in Government at 11, Memo M-25-22 (Apr. 2025), https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf (hereinafter OMB Memo M-25-22).

usually includes common training datasets, agreed-upon evaluation metrics, standard data splits, and published leaderboards. Benchmarks have gained popularity as they provide standardized, scalable, and reproducible metrics across models, teams, and time.[10]

However, conventional benchmarking suffers from several limitations. First, because benchmarks evaluate for static outputs, they are unable to capture real usage conditions, including user interactions, adversarial inputs, or model and data drift, all of which are increasingly important for evaluation of multimodal AI applications in the wild.[11]

Second, benchmarks require context-specific training data and task-specific metrics, which can be difficult to attain for a variety of reasons (e.g. using performance on the bar test as a benchmark proxy for legal reasoning).[12] Training models on decontextualized data and benchmarks raise serious concerns about such benchmarks as measures of task performance.[13]

Third, data leakage, wherein a model is trained on test data, can result in overly-optimistic benchmarking results, but poor performance in real-world deployments.[14] Fourth, benchmarks struggle to keep up with the current pace of AI model development, resulting in issues like performance saturation, that is, when models achieve near-perfect scores on existing benchmarks, eliminating any meaningful differentiation.[15]

Fifth, benchmarks have also traditionally focused on the *technical performance* of an AI system (*e.g.*, accuracy, robustness, and latency metrics), which does not address the full range of considerations important for the increasing contexts where AI is deployed. More recently, technologists have pushed for benchmarking metrics that are crucial for measuring *trust and safety* considerations (e.g., including model bias or disinformation, resilience to adversarial attacks, information security) or broader *program impact* (e.g., whether an AI system has generated measurable value or improved performance). Benchmarks now exist that take a more holistic view of a system's performance, but even such welcome improvements still suffer from the other limitations discussed.[16]

[10] Bernard J. Koch & David Peterson, *From Protoscience to Epistemic Monoculture: How Benchmarking Set the Stage for the Deep Learning Revolution*, arXiv:2404.06647v1 (2024), https://arxiv.org/html/2404.06647v1.

[11] Angelina Wang, Daniel E. Ho & Sanmi Koyejo, *The Inadequacy of Offline Large Language Model Evaluations: A Need to Account for Personalization in Model Behavior*, 6 PATTERNS (2025).

[12] Maria Eriksson et al., *Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation*, arXiv:2502.06559v2 (2025), https://arxiv.org/pdf/2502.06559.

[13] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton & Alex Hanna, *AI and the Everything in the Whole Wide World Benchmark*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (2021), https://arxiv.org/abs/2111.15366.

[14] D. Ramos et al., *Are Large Language Models Memorizing Bug Benchmarks?, in* 2025 IEEE/ACM INT'L WORKSHOP ON LARGE LANGUAGE MODELS FOR CODE (2025).

[15] Shana Lynch, *AI Benchmarks Hit Saturation*, STANFORD UNIV. HUMAN-CENTERED AI (Apr. 3, 2023), https://hai.stanford.edu/news/ai-benchmarks-hit-saturation.

[16] Percy Liang et al., *Holistic Evaluation of Language Models*, TRANSACTIONS ON MACHINE LEARNING RESEARCH (2023).

B. Recent Executive Actions

Within the public sector, recent Executive Orders have begun to impose more systematic AI evaluation requirements, improving on private sector practices. Here we discuss the recommendations of these recent orders, which expand AI evaluation beyond the typical benchmarking approach.

The Biden Administration initiated an expansion of federal AI evaluation requirements with a 2023 Executive Order directing the National Institute of Standards and Technology (NIST) to establish government-wide requirements for developing and deploying AI systems.[17] In 2024, OMB issued directives on government use of AI and procurement of AI.[18] These memos provided guidance to the public sector on several AI-specific evaluation methods, including algorithmic or AI Impact Assessments (AIA), ongoing monitoring, and risk management practices.

Within OMB Memo M-24-10, these required practices applied to any "safety-impacting" and "rights-impacting" AI, namely, any AI system whose outputs provided a principal basis for decisions affecting safety or key individual interests—such as eligibility for benefits, immigration status, or law-enforcement actions.[19]

Under the second Trump Administration, OMB memo M-25-21 explicitly rescinded "rights-impacting" and "safety-impacting" categories[20] and replaced those categories with a single class of "high-impact AI." OMB M-25-10 still retained much of the same evaluation architecture described in M-24-10.[21] First, both memos require agencies to conduct "pre-deployment testing" in environments that "reflect real world" conditions or outcomes, while M-24-10 specified that testing should "follow domain-specific best practices, when available." Second, under both memos, agencies are also required to complete AI impact assessments (AIAs)—formalized risk management tools documenting an AI system's costs and benefits—that state the intended purpose and expected benefit of the AI system, any potential risks, and the quality and appropriateness of the data.[22] Both memos also require an independent reviewer within the

[17] Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Oct. 30, 2023).

[18] Memorandum from Shalanda D. Young, Acting Dir., Office of Mgmt. & Budget, to the Heads of Exec. Dep'ts & Agencies, Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, Memo M-24-10 (Mar. 28 2024), https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf (hereinafter OMB Memo M-24-10); Memorandum from Shalanda D. Young, Acting Dir., Office of Mgmt. & Budget, to the Heads of Exec. Dep'ts & Agencies, Advancing the Responsible Acquisition of Artificial Intelligence in Government, Memo M-24-18 (Sept. 24, 2024), https://www.whitehouse.gov/wp-content/uploads/2024/10/M-24-18-AI-Acquisition-Memorandum.pdf (hereinafter OMB Memo M-24-18)

[19] OMB Memo M-24-10, *supra* note 18 at 5.B

[20] OMB Memo M-24-18 was also rescinded by Memorandum from Mark Paoletta, General Counsel, Office of Mgmt. & Budget, to the Heads of Exec. Dep'ts & Agencies, Memo M-25-18 (Mar. 18, 2025), https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-18-Rescission-of-M-25-17.pdf (hereinafter OMB Memo M-25-18)

[21] OMB Memo M-25-21, *supra* note 8 at § 4.B.I-II.

[22] OMB Memo M-24-10, *supra* note 18 at § 5.C.i.A; OMB Memo M-25-21, *supra* note 8 at § 4.B.II.

agency to identify concerns or gaps in the impact assessment.[23] Third, ongoing monitoring efforts, such as auditing, requires testing for real-world issues with the AI system after its deployment. Both memos recommend human review of AI-enabled decisions, and ongoing procedures to monitor "degradation"[24] or "changes to"[25] an AI system's functionality.

Prior to the Biden Executive Order, seven leading AI companies had voluntarily committed to evaluate generative AI models for societal risks and national security concerns via "red teaming", which was not defined in the voluntary commitments but is generally understood as a structured testing initiative performed by "red teams" using adversarial methods to exploit flaws or vulnerabilities within an AI system in a controlled environment.[26] Following on these commitments, the Biden EO further directed the Secretary of Commerce to require that any developers of covered, high-risk models share the results of red-team testing of these models,[27] and directed NIST to develop further guidance on red-team testing standards for these developers.[28] The corresponding AI acquisition memo, M-24-18, further required that agencies include contractual requirements ensuring that vendors conduct pre-deployment and on-going testing, including red-teaming, and provide the results of such testing to agencies.[29] Yet red-teaming can only surface and resolve so much: one RAND study of LLM-assisted biological attack planning found that LLM assistance did not increase the viability of attack plans relative to an internet-only baseline and that outputs often tracked information already available online.[30]

Recent AI-specific procurement memos published by the Biden and Trump administration offer the most explicit, concrete, and performance-focused techniques for AI governance in current federal law to date. Under the Biden administration, OMB M-24-18 required acquisition teams to assess planned AI procurements and to plan for monitoring and post-award management.[31] The memo also encouraged agencies to prototype high-risk AI systems, test vendor claims, and instructed Chief AI Officers (CAIOs) to collate and share best practices and artifacts for AI acquisition that included guides for "testing, evaluation, and continuous monitoring."[32]

The 2025 Trump administration's memo regarding AI procurement also calls for the development of quality-assurance surveillance plans (QASPs) that define measurable outcomes

[23] OMB Memo M-24-10, *supra* note 18 at § 5.C.iv.C; OMB Memo M-25-21, *supra* note 8 at § 4.B.II.F
[24] OMB Memo M-24-10, *supra* note 18 at § 5.C.iv.D.
[25] OMB Memo M-25-21, *supra* note 8 at § 4.B.III
[26] The White House, *Voluntary AI Commitments* (Sept. 12, 2023), https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.
[27] Exec. Oder No. 14,110, § 4.2.C
[28] *Id.*; Jonathan Spring & Divjot Singh Bawa, *AI Red Teaming: Applying Software TEVV for AI Evaluations*, CYBERSECURITY & INFRASTRUCTURE SEC. AGENCY (Nov. 26, 2024), https://www.cisa.gov/news-events/news/ai-red-teaming-applying-software-tevv-ai-evaluations.
[29] OMB Memo M-24-18, *supra* note 18 at § 4.F.v
[30] Christopher A. Mouton et al., *The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach* (RAND Corp., Rsch. Rep. RR-A2977-1, Oct. 16, 2023).
[31] OMB Memo M-24-18, *supra* note 18 at § 4.C.i
[32] OMB Memo M-24-18, *supra* note 18 at § 3.B.i

and oblige government personnel to "assume a more active role in performance monitoring."[33] These two memos represent the latest thinking in the evolution of AI evaluation methods within the federal government. The focus on continuous monitoring and performance also showcases the shift within AI governance, from technical metrics to trust and safety features, and ultimately the ongoing monitoring of quality and performance.

These mandates have begun to generate concrete institutional responses, though their durability remains in question. Pursuant to the executive mandate for NIST to develop red-teaming standards, the U.S. AI Safety Institute launched the Testing Risks of AI for National Security (TRAINS) taskforce in November 2024, pooling expertise across agencies to conduct joint risk assessments and red-teaming exercises across national security domains.[34] The Department of Energy has partnered with the AI Safety Institute to conduct red teaming using its National Laboratories as AI Testbeds.[35] NIST's Assessing Risks and Impacts of AI (ARIA) program ran its first pilot evaluation, evaluating AI applications using model testing, red teaming, and field testing to assess how systems perform in realistic settings.[36] The pilot demonstrated the feasibility of combining ex-post expert annotation with real-time human tester perceptions to surface risks that neither approach catches alone; for example, annotators and field testers frequently diverged in their assessments of guardrail violations, suggesting that purely

[33] OMB Memo M-25-22, *supra* note 9 at § 4.b.III.B

[34] Press Release, U.S. AI Safety Inst., *U.S. AI Safety Institute Establishes New U.S. Government Taskforce to Collaborate on Research and Testing of AI Models to Manage National Security Capabilities & Risks* (Nov. 20, 2024), https://www.nist.gov/news-events/news/2024/11/us-ai-safety-institute-establishes-new-us-government-taskforce-collaborate. The TRAINS Taskforce was chaired by the AI Safety Institute and included representation from the Departments of Defense, Energy, and Homeland Security, as well as the NSA and NIH. Following the June 2025 rebranding of the AI Safety Institute as the Center for AI Standards and Innovation (CAISI), the operational status of TRAINS is unclear. An August 2025 report recommended that CAISI "leverage" the TRAINS Taskforce, noting that it was "already chaired by the former U.S. AI Safety Institute, offering the opportunity for continuity"—framing that implies continuity had not yet been established. Ctr. for Strategic & Int'l Studies, *Opportunities to Strengthen U.S. Biosecurity from AI-Enabled Bioterrorism* (Aug. 2025), https://www.csis.org/analysis/opportunities-strengthen-us-biosecurity-ai-enabled-bioterrorism-what-policymakers-should.

[35] U.S. DEP'T OF ENERGY, *Artificial Intelligence Testbeds at DOE*, https://www.energy.gov/cet/artificial-intelligence-testbeds-doe. DOE and NIST formalized their collaboration through a Memorandum of Understanding in late 2024, and DOE "activated" pilot AI testbeds at Oak Ridge National Laboratory and Sandia National Laboratories. The DOE testbed infrastructure appears to have greater institutional continuity than the AISI/CAISI programs because it is housed within DOE's national laboratory system rather than depending on Commerce Department organizational decisions.

[36] Razvan Amironesei et al., *Assessing Risks and Impacts of AI (ARIA): ARIA 0.1 Pilot Evaluation Report*, NIST AI 700-2 (Nov. 2025), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.700-2.pdf. The pilot evaluated seven applications from five organizations across three proxy scenarios (TV spoiler shielding as a proxy for privileged information leakage, meal planning for safety risks, and travel planning for hallucination). Testing involved 51 red teamers and 19 field testers and generated over 1,500 expert annotations. Notably, the report was published under the reorganized Commerce Department—Secretary Lutnick appears on its masthead—and two of its eight authors are identified as former NIST employees, underscoring the staffing instability that has accompanied these institutional transitions.

technical evaluation or purely subjective user feedback each misses important dimensions of system behavior.[37]

Yet the institutional trajectory of these promising efforts underscores the structural vulnerability of evaluation mandates resting on executive action. In June 2025, Commerce Secretary Lutnick rebranded the AI Safety Institute as the Center for AI Standards and Innovation (CAISI), explicitly pivoting from safety evaluation toward competitiveness and national security, while earlier mass layoffs at NIST threatened the institute's staffing. CAISI remains active — evaluating foreign AI models, developing security guidelines, and hiring — but its focus has narrowed toward adversary-capability benchmarking and away from the broader sociotechnical evaluation of domestic AI deployments that ARIA was designed to provide. The ARIA pilot report's "Future Directions" section describes plans for subsequent iterations, sector-specific scenarios, and improved measurement tools, but there is not yet public evidence that these plans are being pursued under the reorganized institution.[38] Meanwhile, a January 2025 DHS Inspector General report found that despite appointing a Chief AI Officer and establishing working groups, DHS had failed to develop an implementation plan for its own AI strategy, leaving execution inconsistent across the department.[39]

The broader pattern confirms a structural tension. The 2024 and 2025 memos contain the most concrete requirements anywhere in federal law that agencies (1) articulate hypotheses about how AI systems are supposed to improve performance, (2) specify metrics and data to test those hypotheses, (3) build pre-deployment testing and post-deployment monitoring into governance processes, and (4) embed testing rights into procurement contracts. On the other hand, precisely

[37] *Id.* at 10-14. For one application, field testers perceived greater validity risk (5.00/10) than annotators observed (3.58/10); for another application, the pattern reversed, with annotators recording higher risk (3.52/10) than red teamers perceived (2.24/10). The report treats this divergence as a feature of its multi-method design, noting that "the ability to observe the relative alignment between tester perceptions and expert annotator judgements is a unique advantage of ARIA's approach." This finding is directly relevant to this Article's argument that no single evaluation method—whether technical benchmarking or user feedback—is sufficient on its own. *See infra* Part I.C (discussing limitations of individual evaluation methods); Part II.D (proposing an expanded evaluation toolkit).

[38] The CAISI homepage describes its mission as facilitating "testing and collaborative research related to harnessing and securing the potential of commercial AI systems," with evaluation priorities focused on "demonstrable risks, such as cybersecurity, biosecurity, and chemical weapons" and assessments of "adversary AI systems." NIST*,* *Center for AI Standards and Innovation (CAISI)*, https://www.nist.gov/caisi. CAISI's "Applied Systems" team, which describes its work as leveraging "multidisciplinary methodologies to identify, assess, and measure AI systems in application and real-world settings," may subsume some of ARIA's functions—a December 2025 job posting for the team sought expertise in "rigorous evaluation of end-to-end AI systems in use (e.g., in the workplace and/or in sector-specific tasks)." NIST, *Apply on USAJobs: Open CAISI Position for an AI Research Scientist* (Dec. 19, 2025), https://www.nist.gov/news-events/news/2025/12/apply-usajobs-open-caisi-position-ai-research-scientist. A companion CAISI research blog post similarly frames the work in terms of "AI measurement science" writ large rather than the specific ARIA evaluation methodology. Drew Keller, Ryan Steed, Stevie Bergman & the Applied Systems Team at CAISI, *Accelerating AI Innovation Through Measurement Science*, NIST CAISI Research Blog (Dec. 2, 2025), https://www.nist.gov/blogs/caisi-research-blog/accelerating-ai-innovation-through-measurement-science

[39] Dep't of Homeland Sec., Office of Inspector Gen., *DHS Needs to Do More to Ensure Appropriate Governance of Artificial Intelligence*, OIG-25-10 (Jan. 30, 2025).

because these obligations rest on executive orders and OMB guidance rather than statute, they can be reinterpreted or rescinded with each administration.

C. Limitations of Existing AI Evaluation Methods

Even though the mandates in the executive orders and OMB memos exceed current private sector norms, current AI evaluation methods still face critical challenges for AI system evaluation. In the table below, we provide a taxonomy of evaluation method criteria.

	Typical Stage Conducted	**Measures Causal Effect**	**Temporal Window**	**Assessment Scope**
Benchmarking	Pre-deployment	No	Short	Narrow
Red teaming	Pre-deployment	No	Short	Medium
AI Impact Assessments (AIAs)	Pre-development	No	Medium	Broad
Auditing	Post-deployment	No	Long	Broad

Table 1: Taxonomy of evaluation methods. ***Headers:*** *Typical Stage: when in the system lifecycle the method is typically applied; Causal Effect: whether the method can establish causal relationships between the system and observed outcomes; Temporal Window: the period over which the method gathers evidence and tracks performance; Assessment Scope: the breadth of impacts and effects the method examines; System Outcomes: the scope of what the method can measure.* ***Labels:*** *Pre-deploy: conducted during development and testing before the AI system is released to end users; In-deploy: performed during controlled or staged rollouts where the system is live with real users but under monitored conditions; Post-deploy: executed after full production deployment with real users; Yes: The method uses experimental techniques to establish causal links between the AI intervention and outcomes; No: The method provides observational or descriptive evidence but cannot definitively establish causation due to lack of experimental control; Short: The evaluation captures a snapshot or brief period of system behavior, typically focused on initial performance or pre-deployment assessment; Medium: The evaluation tracks system performance over a transitional period during controlled rollout or early deployment when the system may still evolve; Long: The evaluation provides ongoing monitoring and measurement of system performance over extended periods in production to detect drift, degradation, or changing impacts; Narrow: The evaluation examines isolated components like model outputs or specific vulnerabilities without capturing broader system effects; Medium: The evaluation assesses system-level behavior including decision processes and outcomes in operational contexts, but may not capture full downstream impacts; Broad: The evaluation measures comprehensive real-world impacts including actual benefits, risks, business outcomes, and end-to-end system performance in production.*

The table reveals several points. First, existing methods focus predominantly on evaluating AI systems pre-deployment stage evaluation, which means that these methods could miss important post-deployment issues as they arise, including real-world security vulnerabilities, or model and data drift. Second, current AI evaluation methods provide observational or descriptive evidence of AI system results and outcomes, but cannot draw causal inferences about the effect of the system and intended outcomes. (Drawing a causal inference requires a control group.[40]) For example, benchmarking helps us understand whether an AI model is achieving accuracy scores on diagnostic tasks, but not whether deploying that model causally reduces (or increases) misdiagnosis rates in clinical practice. Third, the methods can typically only evaluate outcomes

[40] *See* Daniel E. Ho & Donald B. Rubin, *Credible Causal Inference for Empirical Legal Studies*, *in* 7 ANN. REV. L. SOC. SCI. 17 (John Hagan ed., 2011).

over a limited period, usually pre-deployment or initial performance, and could miss crucial insights from ongoing monitoring of a system. Auditing, although it can be used to provide ongoing temporal monitoring of a system, is still episodic, and therefore cannot provide systematic information about a system's performance.

Finally, these methods tend to focus on evaluating specific components of an AI system, such as model outputs, predicted risks, or specific vulnerabilities, but are unable to measure system-level outcomes, including context-specific system outcomes, or real-world system impact. For example, auditing might verify that an AI clinical decision support tool is functioning as intended, while benchmarking could demonstrate that the tool achieves high accuracy in identifying patients who would benefit from a particular treatment. However, neither of these methods could tell us whether deploying that tool in a hospital actually improves patient health outcomes, reduces length of stay, increases clinician adherence to protocols, or improves cost-effectiveness of care – the clinical and operational outcomes that ultimately matter. These system-level outcomes depend not just on the tool functioning correctly, but also other important factors such as how the tool integrates with existing clinical workflows, physician compliance with its recommendations.

The limitations in existing AI evaluation methods – which tend to be narrow, static, and component-specific – demonstrates the need to incorporate broader evaluation methods.[41] Recent executive actions, although noteworthy for introducing additional requirements for risk management and ongoing monitoring of AI systems, still miss the mark on ongoing evaluations of a system's overall impact to the extent they rely on these four evaluation methods. In particular, the inclusion of evaluation methods that can establish causality is critical for several reasons. First, such methods allow for performance-oriented assessments of AI integrations, measuring an AI system's outcomes against human baselines rather than abstract standards. Second, they address the thorny challenge of imperfect human baselines – where human performance might itself be biased or error-prone – by avoiding the trap of holding AI to unrealistic standards (such as zero bias) when the relevant comparison should be to existing practices. Third, causal methods address the fundamental epistemic uncertainty about how AI systems will affect outcomes in the real world, moving beyond predictions about what might happen to evidence about what does happen once these systems are deployed.

In the next section, we discuss evaluation methods in existing public law mechanisms which can provide important methods for the AI evaluation toolkit to ensure that evaluations are rigorous, continuous, trustworthy, and meet their intended programmatic and societal outcomes.

[41] *See* Cary Coglianese & Nabil Shaikh, *Management-Based Oversight of the Automated State: Emerging Standards for AI Impact Assessment and Auditing in the Public Sector* 7 (Univ. of Pa. Pub. L. & Legal Theory Rsch. Paper No. 23-45, 2023); Cary Coglianese & Colton R. Crum, *Leashes, Not Guardrails: A Management-Based Approach to Artificial Intelligence Risk Regulation*, RISK ANALYSIS (2025) (describing AI governance as requiring a management-based approach centered on impact assessment and auditing rather than prescriptive or performance-based standards).

II. Recovering Values from Broader Public Law Evaluation Mandates

Federal law already contains multiple frameworks that, at least on paper, require agencies to evaluate their programs, policies, and technology acquisitions. Unlike Part I—which noted AI-specific tools like benchmarking, red-teaming, and audits—this Part focuses on general evaluation mandates that could apply, if they don't already, to the federal government's own use, deployment, and procurement of AI systems. We emphasize three pillars of public law: the Government Performance and Results Act's performance management regime; the Evidence Act's learning agendas and evaluation plans; and procurement and IT investment laws (implemented through the FAR and related statutes) that tie acquisition to testing and performance measurement. By providing context and limitations for each pillar, we show how each framework reflects an aspiration toward systematic evaluation, but each has been only partially implemented and has, so far, done relatively little to structure how agencies evaluate AI. We then revisit the earlier taxonomy of evaluation methods, updated with methods from these frameworks: pilot testing and Randomized Controlled Trials (RCTs), quality assurance (QA), and performance measurement, to show how these methods from public law strengthen and expand the AI evaluation toolkit, both for public and private sectors.

A. Performance Measurement Under GPRA

Enacted in 1993 with broad bipartisan support, the Government Performance and Results Act (GPRA) established what appeared to be straightforward management principles: agencies must set goals, measure results, and report on their performance.[42] Agencies' strategic plans must include "a description of the program evaluations used in establishing or revising general goals and objectives," with program evaluation defined as "an assessment, through objective measurement and systematic analysis, of the manner and extent to which Federal programs achieve intended objectives."[43] One way to view recent OMB guidance on AI impact assessments is as largely operationalizing this same GPRA-style logic—articulating program goals and tracking tailored metrics—in an AI-specific implementation, offering an ongoing, long-term, performance-based framework for AI system evaluation.

Several features of performance measurement, as outlined by GPRA, could align well with AI evaluation. First, performance measurement's focus on multiple indicators towards a performance goal allows for a multi-dimensional evaluation approach. GPRA calls for agencies to take a multi-faceted approach to success, by "establish[ing] a balanced set of performance indicators to be used in measuring or assessing progress toward each performance goal, including, as appropriate, customer service, efficiency, output, and outcome indicator."[44] Instead of a focus only on certain attributes, an AI system's success definition could include not only

[42] Government Performance and Results Act of 1993, Pub. L. No. 103-62, 107 Stat. 285; *see also* THE PROMISE OF EVIDENCE-BASED POLICYMAKING, REPORT OF THE COMM'N ON EVIDENCE-BASED POLICYMAKING (2017), https://www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf.

[43] 5 U.S.C. §306.

[44] 31 U.S.C. § 1115 at (b)(6).

technical performance, or trustworthiness, or user satisfaction scores, but also programmatic impact towards its intended goals. All the indicators would have to be met (and reported on regularly) for the system to "perform" as intended.

Second, GPRA's strong focus on data accuracy and reporting transparency for performance measurement strongly correlates with the need for high data quality in AI systems, particularly in high-stakes environments. In GPRA, agencies are required to provide "a description of how the agency will ensure the accuracy and reliability of the data used to measure progress towards its performance goals."[45] In AI governance, data accuracy could look like high quality data documentation for system trainings and evaluation, regular human review of system inputs and outputs, or ground truth testing. Vendors or agencies could be required to release public reports on an AI system's performance or provide access to independent review of a system's data, decisions, and pipelines.

Lastly, GPRA requires that agencies "provide a basis for comparing actual program results with the established performance goals," namely, establishing a comparison baseline[46] and provide a root cause analysis when performance falls short of the intended goal.[47] Similarly, an AI system must be evaluated against a pre-established baseline, which could be previous system versions, industry benchmarks, or human baselines, and practitioners might be required to analyze and publish documented AI failures, or revise goals if needed.

In principle, GPRA thus recognizes evaluation as a core management function. In practice, its implementation exposes deep tensions in federal performance management. Some agencies used GPRA to support serious strategic planning; others treated it as a compliance exercise.[48] Responsibility for reviewing agency plans fell to OMB just as its management staff was shrinking;[49] agencies faced conflicting statutory demands, with GPRA's push for new data collection colliding with the Paperwork Reduction Act's emphasis on limiting reporting burdens.[50]

Most importantly, GPRA's attempt to link performance metrics and budgets created perverse incentives. As commentators observed, agencies learned to set performance goals they could reliably meet, "ensur[ing] that they can 'pass' their own grading criteria," rather than using metrics to surface problems.[51] Rather than fostering genuine accountability, GPRA often devolved into a paperwork exercise divorced from substantive improvement.

[45] *Id.* at (b)(8).

[46] *Id.* at (b)(7).

[47] 31 U.S.C. § 1116(b)(3).

[48] *See also* Cary Coglianese, *The Limits of Performance-Based Regulation,*50 U. MICH. J. L. REFORM 525 (2017).

[49] Beryl A. Radin, *The Government Performance and Results Act (GPRA): Hydra-Headed Monster or Flexible Management Tool?*, 58 PUB. ADMIN. REV. 307 (1998).

[50] *Id.*

[51] Sidney A. Shapiro & Rena Steinzor, *Capture, Accountability, and Regulatory Metrics*, 86 TEX. L. REV., 1741, 1744 (2008).

The GPRA Modernization Act of 2010 sought to correct these problems by emphasizing the actual use of performance data in decision-making.[52] Yet subsequent GAO surveys found that only a minority of programs had been evaluated in recent years, and managers' use of performance information changed little, even if some "data-use routines" emerged over time.[53]

One specific challenge for GPRA's application to AI governance is that the statute permits agencies to specify performance measures in isolation, rather than requiring them to be assessed in relation to one another. This structure makes it difficult to surface trade-offs that are central to AI-enabled decision-making. The Social Security Administration's most recent GPRA performance plan illustrates the problem. Its first Strategic Goal is to "Optimize the Experience of SSA Customers," with a headline target of reducing initial disability claim processing times to 215 days.[54] The plan repeatedly emphasizes the use of automation to reduce backlogs—an emphasis reinforced by the agency's 2023 AI Use Case Inventory, which lists a new machine-learning tool designed to identify high-risk claims. Yet the performance plan is largely silent on how the quality or accuracy of these automated interventions will be measured. Although the plan includes a separate strategic goal to "Improve the Accuracy and Administration of Our Programs," the associated metrics focus on aggregate over- and under-payment rates, rather than on the accuracy, error patterns, or distributional effects of the automated systems used to accelerate claims processing.

GPRA's focus on regular reporting, multi-dimensional performance indicators, data accuracy and transparency, and analyses of failures can provide useful legal frameworks for AI evaluation. However, the gap between GPRA's theory and its practice serve as an important warning: performance measurement must be seen not just as a compliance exercise, but as an active, ongoing effort to improve an AI system's performance towards larger goals. Agencies could achieve this by linking GPRA-style performance measures to performance-based contracts, with set milestones for vendors to meet at stages within a contract.

B. Evidence-Building Under the Evidence Act

Congress's most ambitious attempt to systematize evaluation across the federal government is the Foundations for Evidence-Based Policymaking Act of 2018 ("Evidence Act").[55] The Act emerged from a 2016 bipartisan commission tasked with studying federal data collection and used.[56] The legislation's central mandate is to require agencies to develop learning agendas and

[52] GPRA Modernization Act of 2010, Pub. L. No. 111-352, 124 Stat. 3866.

[53] U.S. GOV'T ACCOUNTABILITY OFFICE, GAO 13-570, PROGRAM EVALUATION: STRATEGIES TO FACILITATE AGENCIES' USE OF EVALUATION IN PROGRAM MANAGEMENT AND POLICYMAKING 15 (2013), https://www.gao.gov/assets/660/655518.pdf; see also Kristen Underhill, Br*oken Experimentation; Sham Evidence-Based Policy*, 38 YALE L. & POL. REV. 150 (2019).

[54] https://www.ssa.gov/budget/assets/materials/2025/2025BO.pdf#page=16.74. The goal for 2023 had been an average processing time of 164 days. The actual average ended up being 218 days. The agency relaxed its 2025 goal to be 215 days.

[55] Pub. L. 115-435, 132 Stat. 5529 (2019).

[56] Pub. L. 114-140, 130 Stat. 317 (2016).

evaluation plans, also requiring agencies to designate Evaluation Officers, Chief Data Officers, and Statistical Officials to support these efforts. A "learning agenda" is essentially an agency's prioritized set of questions—the key empirical questions it needs answered to improve programs—paired with an evaluation plan that specifies what studies will be done to answer those questions.[57]

Agencies and governments are frequently marketed various AI tools promising improved efficiencies and time savings, but rarely test such vendor claims within context-dependent environments. The Evidence Act provides a framework for two helpful evaluation methods: Rigorous pilot testing in the pre-deployment stage, and RCTs in the early deployment phase of an AI system's launch, would not only offer clear, causal evidence for a system's intended effect, but also surface potential system-wide insights that illuminate needed improvements or design changes.

AI projects in high-impact contexts could be designated as "significant" learning projects requiring evaluation. RCTs require resources and time, and as such, the Evidence Act leaves open the definition of "significant" for agencies to interpret. OMB M-19-23 offers more guidance on designating a "significant" study, including considering the importance of a program or funding stream to the agency mission, the size of the program in terms of funding or people served, or the extent to which the study will fill an important knowledge gap regarding the impact of the program.[58]

Once significance has been established for a project, as per the Evidence Act, submission of an evidence-building plan would be required, including data, methods, analytical approaches, and possible challenges faced in evaluating the project.[59] Further OMB circulars on the implementation of the Evidence Act offer stronger support for these methods. OMB M-20-12 explicitly calls out randomized controlled trials and quasi-experimental designs as the gold standard of impact evaluation.[60]

Federal agencies could look to specific RCT demonstration projects which offer examples of rigorous impact evaluation for high stakes projects. For example, an RCT developed on a collaboration between Stanford RegLab, the Colorado Department of Labor and Employment (CDLE), and the U.S Department of Labor (DOL) to evaluate use of generative AI in benefits administration highlighted several potential trade-offs: although the AI fact-finding assistance

[57] *Learning Agendas*, Evaluation.gov.

[58] Memorandum from Russell T. Vought, Acting Dir., Office of Mgmt. & Budget, to the Heads of Exec. Dep'ts & Agencies, M-19-23, Phase 1 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Learning Agendas, Personnel, and Planning Guidance (July 10, 2019), https://www.whitehouse.gov/wp-content/uploads/2019/07/m-19-23.pdf.

[59] 5 U.S.C § 306.

[60] Memorandum from Russell T. Vought, Acting Dir., Office of Mgmt. & Budget, to the Heads of Exec. Dep'ts & Agencies, M-20-12, Phase 4 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Program Evaluation Standards and Practices (Mar. 10, 2020), https://www.whitehouse.gov/wp-content/uploads/2020/03/M-20-12.pdf (hereinafter OMB Memo M-20-12).

program showed improvements to historical baselines and garnered positive feedback from reviewers, it ultimately did not show an improvement in the initial hypothesis: improved time efficiencies in the control group.[61] The trial further suggests that the system may reduce inter-reviewer variability, a metric that is arguably deeply important given staffing constraints and surging benefits caseloads.[62]

This demonstration project and its insights provide a clear example of the types of learning agendas and evaluation plans outlined in the Evidence Act, particularly important when evaluating AI-assisted decision-making within high-stakes contexts as benefits adjudication. Organizations typically rely on non-randomized user acceptance test results before rolling out a tool; however, the findings of the CDLE trial show that user satisfaction metrics are necessary but insufficient indicators that a tool or system should be purchased or deployed. Rigorous, statistically sound testing of an AI system in the pre-deployment and in-deployment phase will paint a clearer picture of the expected impact and risks while also uncover nuances and trade-offs important for further evaluation and consideration.

Evaluation practice under the Evidence Act has been more modest. Like GPRA, the Evidence Act suffers from a basic resourcing problem. The originating commission explicitly recognized that the "Federal evidence-building community has insufficient resources" and recommended allowing agencies to dedicate a small fraction of program administration budgets to evaluation.[63] The statute did not adopt this recommendation, instead directing agencies to "use existing procedures and systems" to meet new obligations.[64] Unsurprisingly, GAO's early reviews found only "some progress," with agencies citing lack of staff, methodological capacity, and independence as barriers to high-quality evaluations.[65] Treasury exemplifies the resource problem: despite employing over 100,000 staff, the department has fewer than five dedicated full-time evaluation personnel after funding requests for a central evidence coordination team were rejected.[66]

[61] Olivia H. Martin, Varun Magesh, Faiz Surani, Kit Rodolfa & Daniel E. Ho, *Evaluating Generative AI in Benefits Administration: A Demonstration Project*, 5TH ACM SYMPOSIUM ON COMP. SCI. & L. (forthcoming, 2026).

[62] *Id.*

[63] The Commission also proposed that Congress enable agencies to make transfers "across budget accounts to support multi-departmental evidence-building needs." *See* THE PROMISE, *supra* note 42.

[64] The final legislation also omitted the Commission's centerpiece recommendation to establish a National Secure Data Service to facilitate secure data linkages across agencies, as well as recommendations to better connect state administrative data. *Compare* THE PROMISE *with* Pub. L. 115-435.

[65] U.S. GOV'T ACCOUNTABILITY OFFICE, GAO 24-106982, EVIDENCE-BASED POLICYMAKING: AGENCIES NEED ADDITIONAL GUIDANCE TO ASSESS THEIR CAPACITY (2024), https://www.gao.gov/assets/gao-24-106982.pdf. As GAO notes in the report, one of the agencies' chief complaints was the lack of guidance from OMB on how exactly to conduct these capacity assessments.

[66] *Organization and Functions*, U.S. Treasury, https://home.treasury.gov/about/history/history-overview/organization-and-functions. In a proposal for an Evidence Act "2.0," one advocacy group emphasized the importance of incorporating resource allocations, including the 1-percent set aside and the use of evidence in allocating discretionary funds. RESULTS FOR AMERICA, THE PROMISE OF THE FOUNDATIONS FOR EVIDENCE-BASED POLICYMAKING ACT AND PROPOSED NEXT STEPS, https://results4america.org/wp-content/uploads/2019/09/Evidence-Act-Proposed-Next-Steps-FINAL.pdf.

Another limitation is the inconsistent coverage and quality of agencies' evaluations. The Act gives agencies broad discretion in selecting programs for evaluation, leading to wildly inconsistent implementation. For example, the Department of Labor issued a 24-page evaluation plan with 26 evaluation projects–ranging from studying the impact of pandemic unemployment insurance programs to identifying barriers to retirement benefits for underserved communities–for Fiscal Year 2023-2024.[67] The agency's Fiscal Year 2026 plan is five pages with just two evaluation projects.[68] Some of this variation may stem from the Act's relative latitude given to agencies to select programs for evaluation.[69]

Most strikingly for AI, agencies' evaluation plans are largely silent about AI systems, even as AI deployments proliferate.[70] The Department of Homeland Security reported 26 "rights-impacting" AI applications in 2024, including traveler screening algorithms, real-time translation tools for high-stakes officer interactions, facial recognition systems, and fraud detection text mining.[71] None of these systems—nor the terms "Artificial Intelligence" or "AI"—appear in DHS's 2025 Evaluation Plan.[72] Instead, the plan focuses on programs like incentives for minority-serving institutions to conduct homeland security research.[73]

This pattern extends across government. The most recent evaluation plans from SSA,[74] the Department of Labor,[75] the Department of the Interior,[76] the Department of Health and Human

[67] U.S. DEP'T OF LABOR, EVALUATION PLAN 2023-24, https://www.dol.gov/sites/dolgov/files/evidence/DOL-CEO-FY-2023-2024-Evaluation-Plan.pdf#page=21.07. It is unclear, however, how many of these evaluations were actually completed.
[68] U.S. DEP'T OF LABOR, EVALUATION PLAN 2026, https://www.dol.gov/sites/dolgov/files/evidence/CEO-FY-2026-Evaluation-Plan.pdf.
[69] Pub. L. 115-435, 132 Stat. 5529 (2019).
[70] From our experience, the learning-agenda process itself can dilute evaluative ambition: as agendas expand to accommodate numerous stakeholder priorities, they often default toward low-stakes or easily answerable questions, potentially crowding out focused evaluations of high-impact technologies or policies whose assessment would require tighter scope, clearer counterfactuals and sustained analytic investment. *See, e.g.,* https://www.epa.gov/system/files/documents/2022-03/fy-2022-2026-epa-learning-agenda_0.pdf (illustrating a learning agenda centered on stakeholder engagement, measurement infrastructure, and process while treating causal evaluation designs as largely contingent or aspirational).
[71] Office of the Federal Chief Information Officer, 2024 Federal AI Use Case Inventory (GitHub repository), https://github.com/ombegov/2024-Federal-AI-Use-Case-Inventory.
[72] U.S. DEP'T OF HOMELAND SEC., FY 2025 ANNUAL EVALUATION PLAN, https://www.dhs.gov/sites/default/files/2024-03/2024_0310_dhs_fy2025_annual_evaluation_plan_0.pdf
[73] The proposed evaluation will be conducted through an external contractor for two-years with a research design that indicates no control group. Id.
[74] SOC. SEC. ADMIN., FISCAL YEAR 2025 EVALUATION PLAN, https://www.ssa.gov/data/policy/SSA%20-%20FY%202025%20Evaluation%20Plan%20FINAL.pdf#page=33.05.
[75] DOL 2026 EVALUATION PLAN, supra note 68
[76] U.S. DEP'T OF INTERIOR, FISCAL YEAR 2025 ANNUAL EVALUATION PLAN, https://www.doi.gov/sites/default/files/documents/2024-08/fy-2025-annual-evaluation-plan508.pdf.

Services,[77] the Office of Personnel Management,[78] the Department of Justice,[79] and the U.S. Department of Agriculture[80] make no mention of AI evaluation. The Department of Veterans Affairs' most recent evaluation plan at least asks a question that includes AI, seeking to evaluate the adoption rate of "virtual care and digital health technologies (including AI) within VA," though it only specifies use of a "retrospective observational study" with "stakeholder and expert interviews" to answer this question.[81] The disappointing lack of evaluation of AI use cases suggests that the Evidence Act's most promising tools—learning agendas and coordinated evaluation plans—have, so far, done little to structure AI evaluation.

C. Quality Assurance Under Procurement Law

As discussed earlier, the recent M-memos on procurement provide the most concrete requirements for AI governance, particularly post-deployment. These include vendor requirements for post-award monitoring and quality assurance, including agencies requiring vendors to develop and report results from ongoing quality-assurance surveillance plans. These memos, taken together with the AI-specific executive orders, showcase the drive within the federal government for agencies to take an active monitoring role in real-world AI performance. However, beyond the M-memos, the Federal Acquisition Regulation (FAR) and federal IT governance law already provide scaffolding for quality assurance evaluation practices in AI.[82]

Quality assurance is an ongoing, post-deployment monitoring process that, in AI, could detect real-world issues such as model drift or security vulnerabilities, or uncover lab-to-real world validation gaps.[83] The long-term nature of quality assurance offers holistic, system-wide information about an AI system. Several features of the Federal Acquisition Regulation (FAR), which govern agency acquisitions of goods and services and seeks to ensure uniform processing and contracting provisions across federal acquisition,[84] and related statutes thus provide a legal

[77] U.S. DEP'T OF HEALTH & HUMAN SVCS., FY 2026 HHS EVALUATION PLAN, https://aspe.hhs.gov/sites/default/files/documents/7247718a73e9243c8632d8c6666df0e7/HHS%20FY%202026%20Evaluation%20Plan.pdf.

[78] U.S. OFF. OF PERS. MGM'T, FY 25 ANNUAL EVALUATION PLAN, https://www.opm.gov/about-us/reports-publications/fy2025-annual-evaluation-plan.pdf.

[79] U.S. DEP'T OF JUST., FY 2026 ANNUAL EVALUATION PLAN, https://www.justice.gov/media/1404061/dl.

[80] U.S. DEP'T OF AGR., ANNUAL EVALUATION PLAN FISCAL YEAR 2025, https://www.usda.gov/sites/default/files/documents/usda-fy2025-evaluation-plan.pdf.

[81] U.S. DEP'T OF VET. AFF'S, FY 2026 ANNUAL EVALUATION PLAN, https://department.va.gov/wp-content/uploads/2025/06/FY-2026-Annual-Evaluation-Plan.pdf.

[82] *See* Cary Coglianese, *Procurement and Artificial Intelligence*, *in* HANDBOOK ON PUBLIC POLICY & AI (Regine Paul, Jennifer Cobbe & Emma Carmel eds. 2024) (arguing that procurement contracts constitute a first line of defense against improper public sector AI use).

[83] Andrei Paleyes, Raoul-Gabriel Urma & Neil D. Lawrence, *Challenges in Deploying Machine Learning: A Survey of Case Studies*, 55 ACM COMPUTING SURVEYS at 13 (2022).

[84] *See* Erika K. Lunder, Michelle D. Christensen, and L. Elaine Halchin, *The Federal Acquisition Regulation (FAR): Answers to Frequently Asked Questions*, CRS 4–7, 11 (Dec. 18, 2025), https://crsreports.congress.gov/product/pdf/R/R42826; Christopher R. Yukins, *The U.S. Federal Procurement System: An Introduction,* GWU Law School Public Law Research Paper No. 2017-75 (Nov. 1, 2017), https://ssrn.com/abstract=3063559.

foothold for quality assurance: setting pre-determined quality standards for a project, evaluating a system's outputs towards said standards, and ongoing interventions to hold constant the quality of the system overall.[85]

The FAR's almost 2,000 pages includes numerous provisions aimed at encouraging pilot deployments and pre-acquisition prototyping, ensuring agency acquisitions have clearly defined performance standards, and contract requirements for vendors to conduct ongoing monitoring towards those standards. FAR Part 7 requires written acquisition plans to "describe the test program" for each major phase of system acquisition, ensuring agencies plan formal pilots before full implementation,[86] while FAR Subpart 9.3 allows agencies to require "first article testing" when acquiring new or unproven products–which AI systems almost always are–allowing agencies to require initial prototypes for testing before committing to a full purchase.[87] More broadly, FAR Part 46 requires agencies conduct testing to ensure "supplies or services (including commercial services) tendered by contractors meet contract requirements."[88]

For standard setting, FAR 37.601 states that agencies should set, for an acquisition of services, performance standards (and appropriate performance incentives),[89] while FAR 37.604 requires a QASP to be developed either by the agency or vendor.[90] Specifically for IT contracting, FAR Part 39 encourages agencies to apply "continuous collection and evaluation of risk-based data" and "prototyping prior to implementation" for high-risk IT, such as AI.[91] This Part also mandates modular contracting "to the maximum extent practicable," directly supporting pilot deployments and phased testing of AI capabilities.[92] Furthermore, FAR 46.105 holds contractors responsible for "maintaining substantiating evidence . . . that the supplies or services conform to contract quality requirements, and furnishing such information" to agencies as required.[93]

These items, in the context of AI evaluation, indicate several responsibilities for AI vendors. First, they must adhere to, and potentially co-develop, performance standards for an AI system

[85] *See* Daniel E. Ho, Olivia H. Martin, Amy Perez & Kit Rodolfa, *Evaluation as Due Process: Civil Service in an Automated Age*, 78 ADMIN. L. REV. 831 (2025). Agencies are also obligated beyond procurement law by procedural due process principles to set clear quality standards, particularly when individual rights may be impacted by AI systems. Agencies must set administrative processes to ensure a "tolerable average level of accuracy" for decisional quality, as well as ongoing monitoring of said levels, which could include error analyses and bias testing to reduce decision variability. *See* Richard H. Fallon, Jr., *Some Confusions About Due Process, Judicial Review, and Constitutional Remedies*, 93 COLUM. L. REV. 309, 336-37 (1993). Unlike the M-memos, due process and the FAR were not written with AI systems in mind; however, these laws offer important frameworks in line with quality assurance methods: setting quality standards as part of due process, ongoing monitoring via quality assurance surveillance plans, and AI vendor responsibilities.

[86] FAR Subpart 7.015.

[87] FAR Subpart 9.302-9.303.

[88] FAR Subpart 46.102.

[89] FAR Subpart 37.601.

[90] *Id.*

[91] FAR Subpart 39.102.

[92] FAR Subpart 39.103.

[93] FAR Subpart 46.105.

along with purchasing agencies. Second, vendors must develop a plan for and provide ongoing quality surveillance of a supplied system, as well as be ready to provide substantiating evidence of consistent quality standards to a purchaser as needed. Thus, FAR puts the burden of proof on AI vendors, instead of the purchasing agency – an important distinction given agencies' resourcing constraints.

Beyond the FAR, federal IT governance law creates a parallel set of expectations about assessing major technology investments, though some of these operate more through budget and planning channels than explicit procurement law. The Federal Information Technology Acquisition Reform Act (FITARA) requires non-Defense agencies to obtain approval by their Chief Information Officer (CIO) before entering any IT contract.[94] For high-risk IT investments, CIOs must identify the causes of risk, the extent to which these causes can be addressed, and the "probability of future success."[95]

A more longstanding requirement is the Clinger-Cohen Act of 1996's mandate that agencies engage in capital planning and performance measurement for all major IT investments, a requirement implemented through OMB Circular A-11's "IT Business Case" process.[96] Agencies must annually justify all major IT investments by describing expected benefits, documenting life-cycle costs, identifying risks, and providing performance metrics. On paper, this framework appears to impose something like an evaluation requirement on large systems, including those that incorporate AI, because agencies must specify how performance will be measured and how risks will be managed over time.

In practice, however, the "evaluations" included in these business cases focus on whatever metrics are deemed relevant by the agency for that project, sometimes resulting in rather narrow and operational assessments. Take, for example, the Department of Homeland Security's business case for the CBP's automated targeting system (ATS) maintenance.[97] ATS is a predictive tool to target "individuals and cargo that are potential threats." DHS reports several; performance measures, including uptime percentage of the system, the number of shipment exams conducted each year, cost savings of decommissions old legacy infrastructure, the dollar amount of narcotics captured through the system, and the number of travelers denied boarding pre-departure. The DHS reports being "over target" on every single measure of performance. Yet nowhere in these performance measures is any metric of the accuracy of the system or false positives, nor is there any measure relating to timeliness. These indicators satisfy the Circular A-

[94] 40 U.S.C. § 11319.

[95] *Id.*

[96] Clinger–Cohen Act, Pub. L. No. 104-106, 110 Stat. 642 (1996); Office of Mgmt. & Budget, Circular A-11: Preparation, Submission, and Execution of the Budget, https://www.whitehouse.gov/wp-content/uploads/2025/08/a11.pdf.

[97] U.S. Customs & Border Prot., Automated Targeting System (ATS) Maintenance: IT Investment Business Case (Unique Inv. Identifier 024-000005052), https://www.itdashboard.gov/document-search-results/https~3A~2F~2Fogp-s3-prod.s3.amazonaws.com~2Fdata~2Fdocs~2Fbusiness-case~2F2020~2F024-000005052.pdf.

11 requirement to track performance but bear little resemblance to the kinds of testing, validation, or outcome measurement that policymakers often have in mind when discussing evaluation of AI tools.

In short, procurement and IT investment law are a natural legal home for AI testing and evaluation, but current practice often treats their evaluation hooks as flexible, easily satisfied paperwork requirements.

D. An Expanded AI Evaluation Toolkit with Values from Public Law

The promise is there: agencies have clear authority to demand prototypes, pilots, and performance metrics that speak directly to AI risks. Realizing that promise, however, requires reinterpreting and enforcing these frameworks with AI systems squarely in view. We revisit the earlier taxonomy of evaluation methods, updated with methods from public law and discuss further the benefits and limitations provided by the following additional methods, namely, pilot testing and RCTs, quality assurance, and performance measurement.

As shown in Table 2, adding these evaluation methods offer a more comprehensive and holistic evaluation toolkit. In particular, the toolkit's evaluation temporal window is much expanded, allowing for long-term, post-deployment evaluation of an AI system which enables detection of real-world model drift or output degradation. Importantly, pilot testing and RCTs as a method also allows for establishing a causal relationship between the AI intervention and an intended outcome–a crucial method for agencies looking to test vendor claims about a product within their specific context.

	Typical Stage	**Causal effect**	**Temporal Window**	**Assessment Scope**
Benchmarking	Pre-deployment	No	Short	Narrow
Red-teaming	Pre-deployment	No	Short	Medium
AI Impact Assessments (AIAs)	Pre-development	No	Medium	Broad
Auditing	Post-deployment	No	Long	Broad
Pilot testing and RCTs	In-deployment	Yes	Medium	Broad
Quality Assurance (QA)	Post-deployment	No[98]	Long	Medium
Performance measurement	Post-deployment	No	Long	Medium

Table 2*:* Taxonomy of evaluation methods, updated with methods from public law.[99]

[98] Some scholars have argued for mix-and-match interventions which could allow for establishing causality of an AI system with methods such as QA, in which human reviewers evaluate random hold-out sets to understand the impact of AI assistance on treatment sets. *See* David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. REG. 800 (2020).

[99] Headers and labels are defined as in **Table 1.**

By drawing out the benefits and drawbacks of each method across specific dimensions, we also hope to allow practitioners to consider mixing and matching evaluation methods to maximize benefits of specific methods. For example, QA can be matched with other evaluation mechanisms to establish causality, for example, incorporating random hold out sets to test the impact of a new AI feature,[100] while performance measures could be evaluated in shorter temporal windows to match the scale at which AI interventions are happening.[101]

Public law – the GPRA, Evidence Act, due process, and procurement law – all offer useful insights and frameworks for AI evaluation methods. Beyond existing methods, pilot testing, RCTs, quality assurance, and performance measurement offer long-term, ongoing, performance-based evaluations of AI systems, adding important dimensions to the existing AI evaluation toolkit. Taken together, these methods offer a three-tiered risk-based evaluation framework with relevant levels of evaluation rigor. Performance measurement, a relatively low-investment evaluation method, can be required under the GPRA for all AI projects, with clear performance-based metrics and indicators that are reported annually.

Quality assurance, via federal procurement law, requires more active, ongoing monitoring via quality assurance surveillance plans – but can put the onus on AI vendors as part of a performance-based procurement plan. Taken with the most recent federal guidance on AI acquisition, agency contracts could also require vendors to provide data and access for independent verification of testing results.[102] Prior research demonstrates that independent third party audits of results are more truthful[103] – thus, agencies could further consider mandating AI vendors to engage third party evaluators to design and implement their quality assurance surveillance plans. Finally, for high-stakes, high-priority projects, federal agencies can conduct impact evaluation using rigorous pilot testing or RCTs, similar to the CDLE demonstration project.

Mixing evaluation methods together into a holistic evaluation plan, as outlined in the OMB memos, can also help to boost benefits and circumvent cons of single methods. A risk-tailored, holistic evaluation toolkit is necessary for federal agencies to keep up with the rapid proliferation of AI-native applications and systems, both within and outside of the public sector.

[100] Engstrom and Ho, *supra* note 98 at 849.

[101] This typology has an analogue in the literature on management techniques to improve quality and hybrid approaches that borrow elements from different types. *See* Daniel E. Ho & Sam Sherman, *Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement*, 13 ANN. REV. L. & SOC. SCI. 251, 254-55 (2017).

[102] OMB Memo M-25-22, *supra* note 9 at 10.

[103] *See* Esther Duflo, Michael Greenstone, Rohini Pande & Nicholas Ryan, *Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India*, 128 Q.J. ECON. 1499, 1539–40 (2013); Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg & Daniel E. Ho, *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance*, *in* PROCEEDINGS OF THE 2022 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 557 (2022).

III. Paths Forward

The central question in contemporary AI governance is whether AI systems add measurable value that outweighs risks in the settings where they are deployed. That is fundamentally a question about evaluation. Seen through the broader lens of federal performance measurement, evidence-building, and procurement law, two design challenges dominate: ensuring that evaluation is independent of the actors who build and operate AI systems and resourcing that evaluation at a scale commensurate with the stakes. The frameworks surveyed in this Article supply building blocks, but building a governance regime that reliably distinguishes hype from reality will require institutional models that solve, explicitly, for independence and resourcing across the ecosystem.[104] Researchers, funders, governments, and private deployers all have roles to play.

A. Machine Learning Researchers: Treat Evaluation as a First-Class Object

For machine learning researchers, the core challenge is to move beyond generic benchmarks toward context-specific evaluation standards that reflect programmatic goals and legal obligations. That means designing test suites that incorporate domain-relevant outcomes, distributional effects, and error costs rather than optimizing solely for aggregate accuracy on static datasets. Evaluation has lagged behind increasing model complexity, with many systems still being judged on short-horizon tasks even as real deployments increasingly depend on long-context reasoning and robustness in messy, longitudinal settings.[105] It also means revisiting older concepts like quality assurance and performance measurement and translating them into AI-native practices—developing methods for continuous monitoring, drift detection, and field validation that can be integrated into agency QA plans and Evidence Act evaluation designs. Researchers can help institutionalize a modern analogue of Independent Verification and Validation by treating "does this deployed system do what the program says it does, under real-world conditions?" as a research contribution in its own right, not an afterthought.

B. Funders: Resource an AI Evaluation Infrastructure

Foundations and other funders are uniquely positioned to experiment with institutional forms that are hard for governments or firms to build on their own, particularly in areas where evaluation requires independence, technical depth, and sustained investment. One promising path is to draw on the institutional template established by MDRC (then the-Manpower

[104] *See* John Fabian Witt, *How to Save the American Experiment*, N.Y. TIMES (Oct. 6, 2025), https://www.nytimes.com/2025/10/06/opinion/politics/how-to-save-the-american-experiment.html. Witt's historical account of the 1920s democratic crisis emphasizes that recovery depended on the construction of new institutional forms—such as industrial unions and philanthropic research infrastructures—capable of producing trustworthy information, monitoring performance, and correction failure in real time. The parallel for AI governance, we argue, lies in treating evaluation as a core democratic institution.

[105] *See* Sanmi Koyejo, Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on HER Data, NeurIPS 2024 Keynote (2024) https://neurips.cc/virtual/2024/108513 (making this point in the context of healthcare prediction tasks); *see also* Anka Reuel, Ben Bucknall et al., *Open Problems in Technical AI Governance*, TRANSACTIONS ON MACHINE LEARNING RESEARCH (2025).

Demonstration Research Corporation), an independent nonprofit research organization created in the 1970s with major support from the Ford Foundation and federal agencies to rigorously evaluate large-scale social programs using randomized trials and long-run outcome tracking.[106] MDRC's core function was the professionalization of evaluation itself, based on the conviction that "that new social programs, like new drugs, should undergo careful testing to validate their effectiveness."[107] An analogous "MDRC for AI" could play a similar role today, developing and institutionalizing evaluation standards that bridge machine learning and program evaluation, including benchmarking, user-centered testing, sandbox pilots, phased rollouts, and ongoing quality assurance tailored to the risk profile and institutional context of each deployment.

Philanthropy can underwrite evaluation designs that agencies (or private actors) struggle to fund through operating budgets or procurement channels. This includes randomized controlled trials of AI-enabled tools, longitudinal tracking of downstream outcomes, and reviews that assess how systems behave in real institutional settings over time. A dedicated "AI Evaluation Fund," structured as a spend-down vehicle, could seed these efforts, support durable collaborations between agencies and external researchers, and absorb the upfront costs of rapid, iterative evaluation cycles that conventional program evaluation timelines cannot accommodate. The goal is to build an institutional ecosystem whose job is to answer a deceptively simple set of questions: does this system work, for whom, and at what cost?

C. Governments: Mandate Independent Evaluation

For governments, the task is threefold. First, agencies should use existing legal authorities aggressively to require independent evaluation.[108] Evidence Act learning agendas should treat high-impact AI systems as presumptively "significant" evaluation targets; GPRA performance measures should be reoriented to include explicit metrics for AI-enabled components of programs; and procurement and IT investment rules should be interpreted to require pilot testing, IV&V-style third-party review, and ongoing performance monitoring for rights- and safety-impacting AI.

Second, governments must resource that evaluation. The Intergovernmental Personnel Act offers one lever: agencies can bring in academic and technical experts on temporary assignment to design evaluations, serve as technical advisors on AI procurements, and act as informed intermediaries between program offices and vendors. Used well, IPAs and similar mechanisms can provide the kind of "special hires" that earlier generations of agencies relied on to manage

[106] Celebrating 50 Years of MDRC, MDRC, https://www.mdrc.org/about/50-years-mdrc.

[107] *Id.*

[108] Of course, evaluation mandates implicitly require a data infrastructure sufficient to meaningfully measure outcomes. In many agencies, the binding constraint may not be the legal authority to evaluate, but the absence of linked administrative data, standardized outcome definitions, and other infrastructure to track downstream effects. *See* U.S. GOV'T ACCOUNTABILITY OFF., GAO-24-106982, EVIDENCE-BASED POLICYMAKING: AGENCIES NEED ADDITIONAL GUIDANCE TO ASSESS THEIR EVIDENCE-BUILDING CAPACITY 13-14, 17-18 (2024), https://www.gao.gov/assets/gao-24-106982.pdf (noting challenges agencies face in collecting and assessing data to conduct Evidence Act evaluations).

complex, technical procurements, without requiring agencies to build large permanent evaluation staffs overnight. GSA's Office of Evaluation Sciences (OES) offers a concrete template: an applied evaluation unit that supports agencies government-wide by designing and running rigorous evaluations using administrative data. OES has completed over 120 cross-agency evaluations since 2015, and has staffed key roles through IPA-based fellows and other temporary experts, showing how a small central shop can scale evaluation capacity across government.[109]

Third, lawmakers should consider placing guardrails on contractual terms that cut off evaluation. In the public sector, it is hard to justify AI contracts that bar agencies from testing, red-teaming, or sharing evaluation results with oversight bodies. Federal procurement already assumes that the government can inspect and test what it buys.[110] OMB's most recent AI acquisition guidance pushes further, requiring contracts to support ongoing monitoring and independent evaluation on a regular cadence and explicitly forbidding contractual terms that prevent agencies from "internally disclosing" the vendor's testing procedures or the results of testing.[111] Against that backdrop, "no testing," "no benchmarking," or "no red-teaming" provisions in rights- and safety-impacting AI contracts defeat the government's ordinary inspection-and-acceptance posture and frustrate oversight. Congress and OMB could make explicit that "no testing" or "no benchmarking" clauses in rights- and safety-impacting AI contracts are contrary to public policy and unenforceable, much as certain indemnification and nondisclosure terms already are.[112] One place OMB could go further is clarifying what counts as "internal" disclosure. Because M-25-22 simultaneously requires "independent evaluation" and vendor-provided access, agencies should be understood to retain authority to share testing procedures and results with agency-retained independent evaluators (including support contractors or FFRDC-style partners) under appropriate safeguards.

D. Private Vendors: Embrace Evaluation as a Competitive Advantage

Large technology firms have already gone through one revolution in evaluation. In the past two decades, leaders in tech economics and empirical methods have helped normalize a culture in which even small design tweaks—button colors, ranking changes, copy edits—rarely ship without an experiment. Product teams hire economists and data scientists, wire randomized experiments into their infrastructure, and routinely treat deployment decisions as questions that

[109] https://oes.gsa.gov/work/; *see also* Isaac Cui, Daniel E. Ho, Olivia Martin & Anne Joseph O'Connell, *Governing by Assignment*, 173 U. PA. L. REV. 157 (2024). Unfortunately, under the second Trump Administration, OES no longer exists. *See* https://www.gsa.gov/about-us/organization/leadership-directory.
[110] *See* FAR Subpart 52.212-4 (reserving the right to "inspect or test" supplies or services tendered for acceptance).
[111] OMB Memo M-25-22, *supra* note 9 at 10.
[112] *See, e.g.,* FAR Subpart 32.705 (holding that many license agreements contain indemnification clauses that are "inconsistent with Federal law and unenforceable"); Online Terms of Service Agreements with Open-Ended Indemnification Clauses Under the Anti-Deficiency Act, 36 OP. O.L.C. 112 (2012); FAR Subpart 52.203-19 (prohibiting internal confidential agreements in contracts that restrict lawful reporting of waste and fraud to agency investigators).

deserve causal evidence. That infrastructure exists because firms concluded that rigorous evaluation was essential to understanding the value-add of incremental product changes.

Curiously, the use of AI appears to have occupied a different status in many instances. High-stakes systems are often deployed on the strength of internal benchmarks, anecdotal success stories, and qualitative demos rather than disciplined field experiments or independent validation. Private deployers can close this gap by treating model deployment the way they already treat user-facing features: as a decision that should rest on structured testing against clear outcome metrics. That means building or renting the capacity to run experiments on AI-enabled workflows, aligning evaluation metrics with the underlying business or service objective, and subjecting models to third-party checks when they affect rights- or safety-impacting decisions.

Contracts are a critical lever for shaping these incentives. Buyers should refuse "no testing" provisions, insist on clauses that expressly permit experimentation and external evaluation, and treat access to evaluation artifacts—test suites, error analyses, quality-assurance plans—as part of the product rather than a discretionary add-on. Vendors, in turn, can differentiate themselves by embracing independent verification and validation for their systems and by publishing credible evidence of model performance in realistic settings. A market that expects the same level of disciplined evaluation for AI models that it already demands for minor product features will make it much harder for hype to masquerade as value.

* * * *

Taken together, these steps sketch a path toward an AI governance regime in which evaluation is the connective tissue linking research, practice, and law. Researchers would design methods that speak directly to programmatic and legal questions; funders would build institutions whose mission is to deploy those methods in the public interest; governments would use their legal authorities to demand and pay for independent evaluation; and private actors would compete, in part, on the strength of their evidence. The historical lesson from earlier eras of institutional experimentation is that sustained investment in evaluation infrastructure can change what is politically and administratively possible.

IV. Conclusion

The Article opens with a law school nearly barred from testing the AI system it sought to adopt, a small contractual detail that reveals how easily evaluation can be sidelined, even within institutions well positioned to provide it. But a different path was available. Had the prohibition on testing been rejected at the outset, the school could have functioned as a site of rigorous evaluation—benchmarking the system, examining its behavior in real educational and clinical settings, and sharing those findings with the broader policy community.

That alternative world is still within reach for the broader AI ecosystem, though centering AI governance around evaluation naturally raises objections. First, one might object that building evaluation into adoption will slow innovation. That concern reflects a real risk of proceduralism for its own sake. Yet evaluation can also be innovation-forcing: it shifts competition away from marketing claims and toward demonstrated performance in the contexts where systems will be used. It also may lower the cost of iteration: when institutions define measurable objectives, they reduce uncertainty about what works and expand the market for systems that deliver. The result could be faster diffusion of genuinely useful tools, in addition to earlier detection of failure modes that are otherwise discovered only after deployment.

A second objection is that the Article's evaluation commitments may fit less well for agentic AI and other systems whose behavior is dynamic and more autonomous. Agentic systems do make evaluation harder, because performance depends on end-to-end workflows, shifting environments, and interactions with users and downstream systems. But that is precisely why this Article centers lifecycle, system-level evaluation rather than one-shot model assessment. In agentic settings, the governance question becomes whether the overall workflow stays within specified tolerances—accuracy, safety, escalation, recovery, etc.—under realistic operating conditions and over time. The same legal and institutional levers discussed throughout apply with even greater force: pilot testing that mirrors deployment conditions, procurement-backed quality assurance surveillance with access to relevant artifacts and logs, and GPRA/Evidence-Act style performance measure that track downstream outcomes. As autonomy increases, continuous monitoring becomes a basic condition of control.

A third objection is institutional. GPRA, Evidence Act processes, and procurement oversight carry a reputation for performative metrics and, at times, compliance theater.[113] According to this view, building AI evaluation on those foundations invites a familiar charade. This critique should be taken seriously. The point, though, is that these legal mechanisms are scaffolding: they supply durable hooks for measurement, learning, budgeting, and contractual leverage. Whether they become theater depends on design choices that are already within reach—specificity about outcomes and error costs, independent access to data and systems, enforceable audit and testing rights, and credible consequences when vendors refuse evaluation or systems fail to meet contractual standards. These same tools can also be refined to match AI's distinctive features, including post-deployment monitoring, periodic revalidation, and clearer rules around third-party evaluation and disclosure.

[113] *See, e.g.*, Ho & Engstrom, *supra* note 98, at 803 ("When agency administrators can define and game performance measures and lack clear baselines for judging gains from technology adoptions, new systems can erode accountability and foil oversight rather than promote regulatory goals"); Jerry L. Mashaw, *Reinventing Government and Regulatory Reform: Studies in the Neglect and Abuse of Administrative Law*, 57 U. PITT. L. REV. 405, 406 (1995) (critiquing Reinventing Government efforts as "confus[ing] managing with governing"); Sidney A. Shapiro, Rena I. Steinzor & Matthew Shudtz, *Regulatory Dysfunction: How Insufficient Resources, Outdated Laws, and Political Interference Cripple the "Protector Agencies"* at 8 (Ctr. for Progressive Reform, White Paper No. 906, 2009) (describing the implementation of GPRA as "encourag[ing] agencies to develop vague goals [leading] the entire process [to] devolve[] into a meaningless charade"), https://digitalcommons.law.umaryland.edu/fac_pubs/878/.

Our story began with a contract clause and ends with a governance choice. Evaluation can remain the thing that is waived in the fine print, or it can become a routine condition of responsible adoption. Public law has already supplied much of the legal architecture. The work now is institutional but also requires technical work that aligns and enables such evaluation: aligning incentives so that the people who build, buy, and deploy AI systems have durable reasons to measure what matters and to learn from what they find.