Stanford University
Human-Centered
Artificial Intelligence

RegLab

Issue Brief
HAI Policy & Society
Stanford RegLab
November 2026

A Blueprint for Evaluating AI in Public Administration

Danielle Sutton,* Vera Lummis,* Daniel E. Ho

Key Takeaways

Common AI evaluation methods are insufficient to help governments evaluate AI tools in practice. Despite existing legal mandates, government agencies often struggle to implement effective AI evaluation processes.

Agencies already possess "hidden benchmarks" — administrative records of past task performance that can be used to train and evaluate AI models to meet public sector needs. If released, such benchmarks could both measure and drive AI progress.

Agencies should require third-party evaluations as a condition for procurement and conduct randomized AB tests to understand AI impacts as systems are deployed. Collaborating with universities and other research institutions can offer a cost-effective way to draw on independent outside expertise.

Effective post-monitoring deployment will require rethinking traditional quality assurance approaches to include human-AI collaboration on oversight.

The Problem

In June 2026, autonomous AI agents gained unauthorized access to the Australian Medicare statistics portal, marking the first-known instance of an AI agent hacking a government website. This, and other recent cybersecurity incidents, illustrate the propensity of AI systems to function in ways that deviate from intended behavior and the dangers of lack of oversight. They have also inspired renewed policy proposals and legislation on auditing and evaluation. In California, for example, Governor Gavin Newsom signed Executive Order N-9-26 to accelerate the creation of an independent verification organization framework to evaluate AI systems. Private industry has likewise called for embedded evaluators in an effort to pace the frontier. In response to these incidents, Nvidia CEO Jensen Huang urged frontier labs to allocate 80% of their engineering resources to verification and 20% to development, in contrast to industry norms.

AI evaluation is not a new concern; researchers and policy experts have been developing robust approaches to AI evaluation for years. However, recent incidents have brought critical questions surrounding human oversight of AI into mainstream public debate, particularly with regard to high-impact deployments. Nowhere is this concern more pressing than in the government, which is rapidly becoming one of the largest deployers of AI tools in important domains such as healthcare, education, social services, and law enforcement. In 2025 alone, 56 federal agencies reported

* Equal Contribution

3,611 individual AI use cases within government operations, a 69% increase from the previous year.

As government agencies build chatbots and policy drafting tools, and move toward integrating more agentic AI systems into public services, two categories of risk are becoming particularly salient. The first is safety risk: AI tools can be manipulated or attacked, and agentic systems with access to government data and databases can take unanticipated actions with far-reaching consequences. The second is performance risk: AI systems may fail to perform their intended functions, or fail to deliver the intended benefits after deployment.

This brief addresses whether and how agencies can verify the reliability of an AI tool within the highly consequential setting of public administration, where tools are used to deliver public services more effectively and advance agency missions. We outline the limitations of current approaches to AI evaluation in government and present a blueprint for effective, risk-informed evaluation. Finally, we offer recommendations for how the government can play a leading role in shaping standards and practices for evaluating increasingly powerful AI systems in mission-critical areas.

The Gap Between Evaluation Mandates and Practice

Current AI evaluation practices in both private and public sectors are ill-equipped to help us assess how AI tools perform once they are deployed in real-world settings. While benchmarks are widely used to assess the performance of AI systems, they fall short in several key ways. They evaluate a model or tool's responses

Current AI evaluation practices in both private and public sectors are ill-equipped to help us assess how AI tools perform once they are deployed in real-world settings.

based on a fixed set of questions, which may not reflect performance in actual workflows, and they are prone to saturation over time. Common evaluation methods beyond benchmarking, such as red-teaming and AI impact assessments, help identify vulnerabilities and anticipate risks before a tool is adopted. However, they focus predominantly on assessing tools for a short period before deployment. This fails to measure how tools will impact downstream outcomes, such as decision quality, timeliness of service delivery, and equity. These methods also do not compare the results of using AI tools to non-AI alternatives, failing to measure the actual impact of AI.

Inadequate evaluation of AI tools in public administration heightens the risk of unintended or inequitable outcomes, such as compromising access to public services and eroding due process. Between 2013 and 2015, the automated fraud detection system used by Michigan's unemployment insurance agency accused nearly 40,000 applicants of fraud with an error rate of 93%, resulting in the wrongful denial of vital public resources. More recently, an audit of state unemployment insurance chatbots revealed

that generative AI chatbots introduced new risks by presenting inaccurate information with confidence. For example, reports on New York City's generative AI chatbot found that the tool gave faulty legal advice, incorrectly telling landlords that they could reject tenants with Section 8 housing vouchers. Rigorous evaluation is designed to catch this kind of error before it reaches the public. These cases all involved AI tools that were deployed without adequate evaluation and oversight, demonstrating that evaluation is not only good management, but a safeguard for due process. These risks give the government strong incentives to build robust AI evaluations.

Evaluation of government AI is also rooted in law. The Office of Management and Budget (OMB) mandates evaluation beyond conventional benchmarking. Memorandum M-25-2, issued in April 2025, required pre-deployment testing in real-world conditions, AI impact assessments (AIAs), independent reviews of those assessments, and ongoing monitoring. In parallel, Memorandum M-25-22 extended these obligations to procurement, requiring pre-award testing and ongoing evaluations within the contract terms. The Government Performance and Results Act (GPRA), the Evidence Act, and Federal Acquisition Regulation (FAR) reinforce executive mandates by establishing statutory frameworks that provide long-standing requirements for government evaluations. States have similarly begun to adopt their own requirements: California's generative AI guidelines, for example, require agencies to test models adequately before deployment.

Even so, in practice, government agencies struggle to implement robust evaluation procedures. The Department of Health and Human Services reported 447 active AI uses, yet its most recent Evidence Act evaluation plan does not even mention AI. The same is true of the 2025 evaluation plans of the Social Security Administration, the Department of Labor, Department of the Interior, Department of Justice, Department of Agriculture, and the Office of Personnel Management.

Despite existing legal mandates, in practice, government agencies struggle to implement robust evaluation procedures.

Recommendations: Leverage Existing Frameworks to Promote Evaluation

Agencies already have frameworks to ensure robust evaluation within the public sector. While much conversation has focused on proposals for embedded evaluators at frontier AI companies, less attention has been directed to government procurement rules — one of the most direct ways to govern AI use in government. The government is a major buyer of AI, which means the terms of their contracts with developers largely determine what an agency can learn about how well an AI tool works. This is where the core insight behind embedded evaluators — that independent parties should be able to scrutinize AI systems — can be put into practice.

Agencies should require independent, third-party evaluation as a condition of procurement. They can do so using procurement authority under the FAR (or state analogues), along with the evaluation and performance requirements of the Evidence Act and GPRA (or state analogues). Contracts can define "independent" and mandate that vendors grant evaluators access to models, data, documentation, and compute. There is precedent for this approach: Beginning in the 1970s, federal agencies built a vibrant evaluation ecosystem for public services by partnering with independent institutions to rigorously assess social programs.

Colorado offers an example. When a large vendor approached the Colorado Department of Labor and Employment about using AI to train adjudicators in unemployment insurance, the final contract required the vendor to work with a third-party evaluator and to verify 90% accuracy before advancing to later stages in the contract. Stanford RegLab served as the (pro bono) evaluator.

AI evaluation should not be viewed as a static, one-time process. Instead, government agencies should use existing authorities to support evaluations across the entire life cycle of AI deployment, from pre-deployment to in-deployment to post-deployment evaluations.

1. Pre-deployment: Develop Public Administration Benchmarks By Leveraging "Hidden Benchmarks"

Generic, off-the-shelf AI benchmarks fail to capture the operational realities, edge cases, and shifting contexts inherent to public administration. Meanwhile, agencies often already possess much of the data that can form a compelling basis for benchmarks. Every time an employee corrects a draft report, produces a commentary on a decision, or answers a constituency inquiry, an agency generates a potential record of appropriate responses to a given problem in its setting. If released in privacy-protecting ways, these records, or "hidden benchmarks," could attract engineering talent to address public challenges and help agencies and developers produce more rigorous assessments of the promise, limits, and risks of AI tools for actual tasks.

Every time an employee answers a constituency inquiry, an agency generates a potential record of appropriate responses to a given problem in its setting.

To illustrate, consider three common government use cases:

Citizen helplines. Chatbots that serve as public helplines — such as the federal government's America.gov and California's AskCA — represent some of the most widely implemented government uses of AI. They are difficult to evaluate because agencies cannot anticipate the full spectrum of inquiries constituents might submit. To design evaluations that address the broad range of questions and concerns constituents may raise, agencies can leverage existing logs of 311 calls, emails, helpline recordings, and mail, along with quality assurance reviews of responses to constituent

questions, to build a benchmark that tests whether a chatbot responds accurately and safely.

Complaint review. Agencies receive large volumes of complaints that staff must review, categorize, and prioritize. When AI tools are used to automate elements of this triage process, they can fail in ways that compound over time. Records of how staff prioritized and routed complaints could serve as a "hidden benchmark" — transforming operational data into a valuable measurement tool. An example is the Consumer Financial Protection Bureau's Consumer Complaint Database, which maintains a public archive of consumer complaints submitted to the agency and routed to companies for resolution.

Rulemaking. Many agencies have explored using AI to improve the slow rulemaking process. One theory is that if AI could help draft rule proposals, it could streamline the lengthy internal revision process typically required to correct errors made by line-level human rule-writers. And yet, those very revisions offer a unique benchmark for assessing whether AI tools in fact avoid common drafting errors. Similarly, because agencies must respond to public comments when revising a rule, these revision logs provide an untapped source of benchmark data to evaluate an AI system's capacity to address such comments.

If released, such benchmarks could both measure and drive AI progress, while also serving as an important checkpoint for risks. While the private sector has myriad benchmarks to test AI tools, a gap persists for public administration. Publishing "hidden benchmarks" from government agencies would help to address this deficit and would serve as a natural progression of the open data movement by highlighting public challenges for the AI community to solve.

2. In-deployment: Evaluate AI Impacts with AB Testing

Benchmarking goes only so far. If an agency decides to deploy an AI system, it should also conduct AB tests to understand the impact, benefits, and risks of the system. Gathering causal evidence while a system is in deployment is particularly valuable because it can inform decisions about ongoing rollout. Stanford RegLab, the Colorado Department of Labor and Employment (CDLE), and the U.S. Department of Labor demonstrated this approach in a joint project to evaluate generative AI tools for modernizing unemployment insurance. The team ran a randomized controlled trial of an AI fact-finding assistant on real, historical cases and measured both decision quality and fine-grained behavioral data. Although government personnel found the AI assistant helpful, the system did not substantially improve quality or efficiency compared with a control group. This outcome shows the value of using randomized controlled trials to test AI systems in high-stakes contexts, such as benefits adjudication, and how that evidence can guide future deployment.

Such AB tests, or randomized controlled trials (RCTs), are often perceived as resource intensive, but any rollout involves phasing in, training, and sequencing — providing an ideal opportunity for impact assessment. Furthermore, by collaborating with universities and other research institutions, agencies can draw on outside expertise to offset costs. Third parties can help design the trial, define outcome measures, and analyze the results, while program staff contribute the operational knowledge and access to administrative data necessary for a realistic evaluation.

3. Post-deployment: Move Quality Assurance Toward AI Evaluation

Once a tool is deployed at scale, agencies need mechanisms to monitor its performance over time. For example, with the rollout of AskCA, it is critical for agencies to monitor both safety risks — when the chatbot might be misused or manipulated — and performance risks — where and when the chatbot might give inaccurate or inappropriate advice. These questions become more pressing as agencies deploy agentic AI tools that are integrated with internal systems.

For government, effective post-deployment monitoring of AI tools will require rethinking traditional quality assurance approaches. Historically, such systems have relied on random sampling, such as call-center supervisors reviewing a random subset of recorded interactions to check for quality. Given the immense scale of digital services, AI models can be used alongside traditional quality assurance to monitor and pinpoint chatbot interactions most prone to errors or apt to cause harm.

Moving quality assurance toward AI oversight requires a shift in agency organization. Staff who once responded to calls and email directly may increasingly allocate time to oversee AI systems in the pipeline. This would involve reviewing flagged cases, identifying patterns of failure, and deciding when a tool should be paused or modified. Agencies will need to define these new roles and establish clear authority to intervene when a system is failing. Combined with periodic third-party audits, these practices give agencies ongoing visibility into the performance of AI tools and ways to detect when they go astray.

Effective post-deployment monitoring of AI tools will require rethinking traditional quality assurance approaches.

Conclusion

As AI systems become more capable and more deeply embedded in public service delivery, the central question for the government is not only whether a model is safe before release, but whether a tool is safe and creating value throughout deployment. The government is well positioned to lead in answering that question. Agencies possess administrative data (i.e., “hidden benchmarks”) and institutional knowledge that make robust evaluation possible, have the legal mandate to require evaluations, and, as major purchasers of AI systems, can set standards for evaluation that shape practices more widely. Codifying improved evaluation practices within government will set a precedent for both the public and private sectors, ensuring that evaluation focuses not only on the safety of model capabilities, but on their performance and impacts once deployed.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices.

Stanford University's Regulation, Evaluation, and Governance Lab (RegLab) partners with government agencies to design and evaluate programs, policies, and technologies that modernize government.

The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**



Danielle Sutton is a JD candidate at Stanford Law School, focusing on AI governance, and a Law Student Fellow at the Regulation, Evaluation and Governance (RegLab) at Stanford University.



Vera Lummis is a master's candidate in international policy at Stanford University, specializing in AI policy, and a research assistant at Stanford RegLab.



Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science and Computer Science (by courtesy), Associate Director and Senior Fellow at the Stanford HAI, Senior Fellow at the Institute for Economic Policy Research (SIEPR), and director at the RegLab.

Acknowledgments

This work was supported in part by Stanford HAI, Stanford Impact Labs, and Arnold Ventures.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**