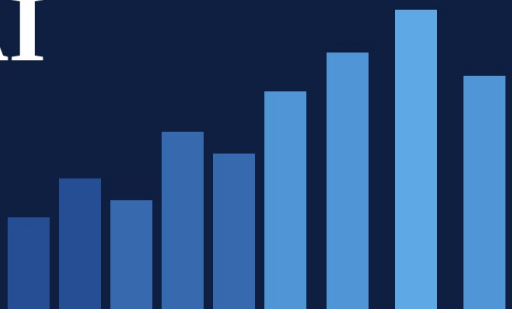
GOVERNMENT AI IN PRACTICE

Research and analysis from the ThinkCapital GIAG Initiative

ISSUE 7 · MAY 20, 2026



EARLY SIGNAL: FROM THE RESEARCH

GIAG Stream Two has begun structured intake interviews with government technology practitioners who hold or have held AI oversight responsibility. An early finding is taking shape: when asked to identify the specific point within any deployed AI workflow where a human decision is required before the system continues, most respondents describe authorization before deployment, or review of outputs after. Often there is no gate inside the running process.

This pattern holds across federal and state agency types and across deployment contexts ranging from benefits eligibility tools to procurement screening systems.

In no early-intake case has a practitioner described a mid-process intervention mechanism that pauses agentic AI execution pending human review. The oversight architecture, in every case examined to date, sits outside the operational envelope of the system it is governing.

FROM THE EDITOR

Government AI oversight requirements are filling policy documents faster than they are shaping deployment architecture. Most agency governance plans describe what oversight should exist at a system level. Most practitioners, when asked how oversight works, describe something structurally different.

The gap matters for a specific reason. When the first significant failure in an agentic AI government deployment occurs (and this will certainly happen) the investigation will ask where the human was. The documentation will show a review process existed. The harder question will be whether the reviewer was positioned to see and stop what went wrong, or whether the review sat downstream from the consequential decisions, looking at outputs after the process had already run. Most current government AI oversight architectures will not survive that question.

This issue describes the problem structurally. It also identifies three actions any government IT leader can take now, before a failure makes the question urgent.

First, map every deployed AI workflow and find the point where a human can pause execution mid-process. If you cannot identify that point, you have documentation of oversight, not functioning oversight.

Second, conduct a scope audit on each deployed system. Compare what the system was authorized to do at initial deployment against what it is currently integrated with and acting on. The gap between those two descriptions is your unreviewed operating perimeter.

Third, define the reviewer's task in operational terms: what information does the reviewer have access to, what specifically are they assessing, and what can they do with a finding. A governance framework that names a reviewer without answering those three questions has not assigned oversight. It has assigned a witness.

~ Michael

THE INTERVENTION POINT PROBLEM

Why Human Review in Government AI Works Differently Than Most Governance Frameworks Assume

Government AI governance documents overwhelmingly require human oversight. While the requirement is genuine, the problem is that “human oversight” is non-specific and subject to interpretation in those documents. The operational detail behind it is thin.

When you ask practitioners to walk through what oversight looks like in their deployments, a consistent picture emerges. Typically, there is a review before deployment (an approval gate, and sometimes a formal risk assessment). There is a review after the system acts, via output audits, periodic reporting, and anomaly flagging. What is missing is structured review of the decisions the system makes in the middle, while it is running. This is the *intervention point problem*.

Where the Model Breaks

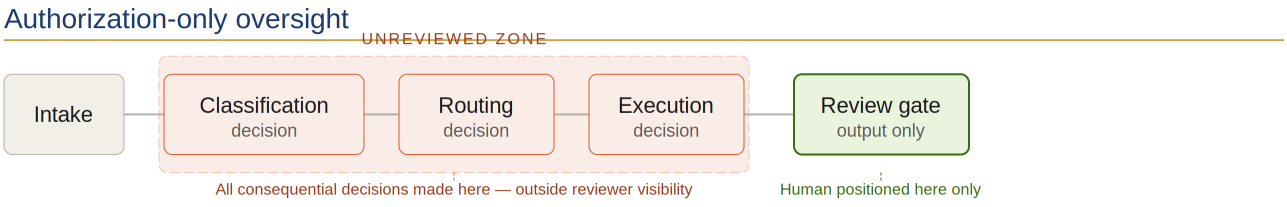
Every government agency deploying AI has adopted, deliberately or by default, one of two oversight models.

Model one: a gate before deployment and a log review after. The system runs between those two endpoints without structured human engagement at any decision point. The approach is to authorize before and audit after.

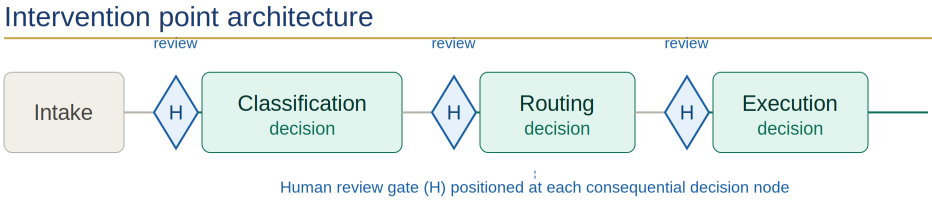
Model two: human review gates positioned at consequential decision nodes inside the process, before the system continues to the next step. Human authorization at the point where choices are being made, not at the point where results arrive.

Empirical findings from Stream Two indicate that more than half of agencies are using the first model. An increasing number have governance documents that describe the second.

When oversight is placed only at the edges of a process chain, consequential decisions accumulate between those endpoints without human authorization. In an agentic system, those intermediate decisions are not advisory outputs awaiting review. They are committed choices that constrain every step that follows. By the time a reviewer sees the result, the decision sequence has already run to completion. Reviewing the output tells you whether the result looks right. It does not tell you whether the process that produced it stayed within appropriate limits.



GIAG Stream Two · ThinkCapital LLC · thinkcapital.org



GIAG Stream Two · ThinkCapital LLC · thinkcapital.org

The Scope Drift Problem

The authorized scope of a deployed AI system and its actual operating scope tend to diverge over time. When oversight sits at the process edges, that divergence is invisible.

Consider how this plays out in practice. A government AI system is authorized at deployment with a defined scope. Six months later, the scope is wider, with more integrations, more data sources, and more task delegation. Each addition looks minor in isolation, and none triggers formal re-examination or re-authorization. In aggregate they produce a system doing significantly more than what was approved. The oversight mechanism, calibrated to the original specification, is checking against a description that no longer matches the deployed system.

Drift is avoided when oversight mechanisms are calibrated against the actual operating scope, not the specification defined at deployment. Most current governance frameworks do not maintain that calibration. They assess scope once, at authorization, and do not revisit it until an audit or an incident.

Four documented cases illustrate the structural pattern. (See companion research note.)

> **Case 1: CBP's Automated Targeting System.** U.S. Customs and Border Protection's Automated Targeting System was authorized as a cargo risk-scoring tool. Its operational scope expanded over time to incorporate passenger travel patterns, financial transaction data, and social media review activity. Privacy impact assessments lagged actual data use expansion. When GAO and DHS OIG examined the system, they found that operational data inputs exceeded what the original authorization documents described. The oversight mechanisms were calibrated to the original cargo-screening scope and had not been updated to reflect the expanded data universe the system was using.
>
> **Case 2: TSA Facial Recognition.** The Transportation Safety Administration Credential Authentication Technology (CAT) program was piloted at specific checkpoints for identity verification. It compares a traveler's face to their identification document photo. By 2024, the program operated at more than 80 airports. A GAO study found that TSA had not completed required privacy studies before expanding the program. The authorized scope (face-to-ID matching at defined checkpoints) had in practice extended to data retention periods and cross-system matching arrangements that the original authorization did not specify, creating legal exposure. Oversight mechanisms designed for the pilot were not redesigned as the deployment scaled.
>
> **Case 3: Arkansas Medicaid Care Algorithm.** The State of Arkansas deployed an algorithm to calculate in-home care hours for Medicaid beneficiaries. The system was authorized based on a defined set of medical conditions. In practice, it incorporated behavioral assessment scores and nurse evaluation factors not in the original specification and not disclosed to recipients. Federal courts found that recipients had no meaningful access to the factors driving their care determinations. The scope of what the algorithm was weighing had expanded beyond its authorization, and the oversight architecture had not tracked the expansion.
>
> **Case 4: VA Claims Processing AI.** The Department of Veterans Affairs deployed AI tools to assist with disability claims processing, initially scoped as decision-support: surfacing relevant records for human reviewers. In practice, several deployed tools took on a more active role in shaping reviewer attention, effectively filtering which evidence received scrutiny. VA Office of Inspector General reviews have noted gaps between the described function of AI tools in claims processing and their operational behavior as deployed. The oversight frameworks designed for an advisory role were not updated when the actual scope of the tools' influence on reviewer decisions became clearer.

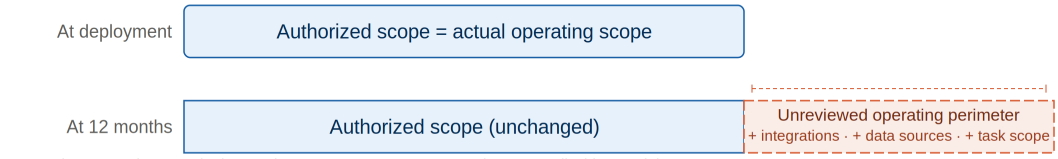
The pattern across all four cases is the same: the scope expanded, and oversight did not. Each expansion was rational in isolation. A more capable system serves users better. An additional data source improves accuracy. A broader deployment reaches more people. The problem is architectural: oversight mechanisms do not scale with operational scope because formal tracking of divergence does not happen.

Gaps between what a system is authorized to do and what it is actually doing are not visible through standard compliance reporting. They require active operational monitoring.

The question for any government IT leader: is your agency tracking the divergence between authorized scope and actual operating scope for each deployed system? If the answer is no, someone is reviewing against the wrong specification.

Full citations for the four cases above are available in the companion research note at thinkcapital.org/research.html.

Scope drift: authorized vs. actual operating perimeter



Pattern documented across cited examples: CBP ATS · TSA CAT-2 · Arkansas Medicaid · VA Claims AI
Each expansion was rational in isolation. The problem is architectural: oversight mechanisms were not updated to track the divergence. GIAG Stream Two · ThinkCapital LLC · thinkcapital.org

What Effective Oversight Requires

Stream Two research is developing a typology of oversight models observed across government AI deployments. The range is wide: some agencies have built genuine intervention point architecture, using review gates at consequential decision nodes inside the process. Most agencies have not and simply review outputs. The modal pattern is authorization-plus-audit: a gate before deployment and a log review after, with no structured visibility into what the system is doing while it runs.

Oversight that functions as a genuine control, rather than a documented assumption of control, requires three things that current governance frameworks routinely underspecify.

- **Intervention points at decision nodes, not output endpoints.** For agencies deploying agentic AI, this means identifying the specific decision nodes within the system's action sequence that require human authorization before the system continues, before deployment. Those gates need to be designed into the architecture, not described in the governance plan.
- **Operationally defined scope, actively monitored.** Agencies need a mechanism for tracking what the system is doing: what data sources it is accessing, what integrations have been added since initial deployment, what tasks it is performing outside the original specification. This requires ongoing monitoring, not a one-time pre-deployment assessment.
- **Reviewers equipped for the actual cognitive task.** Reviewing an agentic AI process for integrity and scope adherence requires access to process-level information (not just outputs) and the domain expertise to evaluate what that information means. Governance frameworks should specify, for each review role, what information the reviewer has access to, what they are specifically evaluating, and what authority they have to act on their findings. A reviewer looking at a summary of outputs from a system they cannot observe in operation is not performing oversight. They are performing sign-off.

All three requirements are concrete enough to evaluate for any specific deployment. Intervention point architecture either exists in the process design or it does not. Scope monitoring either tracks operational divergence, or it does not. Reviewers either hold access to process-level information with defined authority to act on their findings, or they hold a documentation role. These are architectural questions about a specific system, not general policy judgments, and they can be answered with precision.

Stream Two has operationalized this evaluation. The working prototype below applies the five-criterion framework from Working Paper Two to any agentic AI task described in plain language, returning a structured assessment with criterion scores, rationale, and an intervention level recommendation. Running an assessment does not require policy expertise or familiarity with the underlying research. It requires only a brief description of the task.

APPLIED RESEARCH: THE GIAG ASSESSMENT TOOL

The tool takes a plain-language description of any agentic AI task and evaluates it against the five criteria from Working Paper Two: irreversibility, consequence transfer, distributional novelty, value conflict, and legal or regulatory significance. Each criterion is scored independently. The combined profile produces one of five intervention level recommendations: Autonomous, Monitor, Review Required, Joint Execution, or Human Only. The output is a structured Governance Assessment Card with criterion scores, rationale for each score, and suggested oversight actions.

No account required. No login. Three example tasks are preloaded. The tool runs on any device with a browser.

A demonstration video runs two cases back-to-back. The first is a state DMV processing vehicle registration renewals autonomously, with exceptions routed to human staff. Mixed score profile. Recommendation: Review Required. The second is a county social services agency sending automated benefit denial letters without human review of individual cases. Five criteria. Five High scores. Recommendation: Human Only. That discrimination between deployment contexts is what operational AI governance looks like in practice.

Tool: https://www.thinkcapital.org/tools.html

Demonstration video: https://www.thinkcapital.org/Docs/GIAGAssessmentDemo.mp4

Working papers and methodology are found at: https://www.thinkcapital.org/research.html

FROM THE RESEARCH

Stream One:
Working Paper 3 (Mandate Translation) is in active construction and on target for release later this month. The paper examines how federal AI governance frameworks, such as the Office of Management and Budget memorandum M-25-21 defining the Chief AI Officer (CAO) arrive at state and local agencies and what happens to their operational substance in translation.

Preliminary research indicates that most states treat federal AI mandates as compliance signals rather than operational requirements, producing governance documentation aligned with federal standards but oversight architectures that have not been operationally redesigned. Practitioners with state or local AI governance experience are invited to reach out at michael.bragen@thinkcapital.org.

Stream Two:
A scheduled practitioner session in early June on human intervention in government agentic AI deployments will serve as the primary data-collection event for Working Paper 5 (Oversight Typology). If you have direct experience with AI oversight design in a government context and are not already in the research pipeline, the participation form is at https://www.thinkcapital.org/research.html.

FIVE QUESTIONS FOR PRACTITIONERS

These questions come from Stream Two research conversations. They are a starting point for identifying where oversight architecture is thinner than governance documents suggest.

1. Can you identify the specific points in your deployed AI workflows where a human decision is required before the system continues—not a review of output, but a gate that pauses execution mid-process?
2. Has the operating scope of your current AI deployments (what they are connected to and what they are acting on) changed since initial authorization? If so, has that change been formally reviewed?
3. When your reviewers evaluate AI system outputs, do they have visibility into the process steps that produced those outputs, or are they working from results only?
4. Has your agency defined what a reviewer should do when uncertain about an AI-generated output? Is there a documented protocol that defines a method of escalation?
5. How does your governance framework distinguish between oversight of assistive AI (recommendations, analysis) and agentic AI (autonomous action sequences)? Or does it apply a single oversight model to both?

Government AI in Practice

Published weekly by ThinkCapital LLC under the Government IT and AI Governance Initiative (GIAG). GIAG is a practitioner research program examining AI governance implementation in federal, state, and local government. Research participation, practitioner inquiries, and correspondence: michael.bragen@thinkcapital.org. Archive and publications: thinkcapital.org.

WP2 — The Oversight Illusion: When Humans Must Intervene — is available at thinkcapital.org/publications.

The views expressed are those of the researcher. Not for distribution without permission. Michael Bragen, Principal, ThinkCapital LLC | michael.bragen@thinkcapital.org | thinkcapital.org | thinkcapital.substack.com