GOVERNMENT AI IN PRACTICE

Research and analysis from the ThinkCapital GIAG Initiative

ISSUE 10 · JUNE 10, 2026

EARLY SIGNAL: FROM THE RESEARCH

Last week, a working group practitioner session cut directly to what Stream Two research is measuring. The questions are presented here because they highlight key gaps in existing AI governance.

First, when a system causes harm through a sequence of individually approved actions, who owns the outcome? Second, in multi-agent environments, how does an agency govern the combinatorics of agents communicating with each other? Third, should litigation exposure be treated as a fourth governance clock? These questions do not have standard answers in any published framework.

From the Editor

The governance question in most agencies is still binary: do we have oversight? Yes or no? That framing was adequate for advisory AI, but it falls short for agentic AI, where the system is now an actor that allocates, routes, processes, and decides, often faster than any human review cycle can keep up with.

Stream Two research has been building the empirical case for a different question: is oversight present at the moment it must be, and does it stay out of the way when it adds no value, only creating latency? The practitioner evidence from structured interviews and the recent session makes it clear that most agencies have not answered that question for their agentic deployments. This issue presents what the answer requires in operational terms.

~ Michael

When Oversight Requires a Decision: The Determinism Boundary in Practice

Stream Two Findings on Agentic AI Governance Across Federal and State Agencies

FEDERAL

The Determinism Boundary

Federal agencies are crossing a governance threshold without a governance architecture to match it. Deterministic rule-based systems that have been in production for years, processing veterans' benefits, permit decisions, and eligibility determinations, are being migrated to generative AI for speed and capability gains. The governance that covered the old system does not cover the new one, and most agencies have not yet recognized that the crossing requires a deliberate governance transition.

The difference is structural. A deterministic system always produces the same output from the same input. Errors are systematic, findable, and fixable. The accountability chain is clear: someone approved

the rule set. In a generative AI replacement, the same input may produce different output, errors are distributional and harder to find or attribute, and the accountability chain fractures. Someone approved the model, someone approved the prompt, someone approved the deployment, but when something goes wrong no single point owns the outcome.

The governance question that almost no agency can answer before deployment: what governance architecture do you need when the output is no longer deterministic and the consequences are financial, legal, or mission-critical? Stream Two research is documenting what that architecture looks like where it exists, and what its absence produces where it does not.

The Binary Oversight Problem

The Working Group session produced a specific formulation of a finding Stream Two research has been building since January. The governance question in most agencies is still binary: do we have oversight? That framing was adequate for advisory or generative AI where the system is an advisor. It is not adequate for agentic AI, where the system is an actor.

The binary framing produces two specific failure modes. First, oversight that exists but is positioned after the consequential decisions have already been made. A system that processes 10,000 decisions per day and routes each for human review produces a rubber stamp within weeks, documented liability without genuine control, which is worse than no review because it creates false assurance. Second, oversight that is nominally present but lacks the information or the operational authority to act on a finding. A named human with documented authority who cannot access the decision logic that produced the output, or who cannot halt the system without triggering an escalation chain that takes days, does not constitute functioning oversight.

The governance decision that most agencies have not made is which conditions require mandatory human intervention, and which conditions make human oversight counterproductive. Governance that applies uniformly across a portfolio of deployed systems with different consequence profiles is not rigorous governance, it is uniform process application. The distinction determines whether oversight produces control or produces documentation.

STREAM TWO: THE FIVE INTERVENTION CRITERIA

Stream Two research identified five characteristics that consistently require mandatory human intervention, regardless of volume, speed, or efficiency arguments. If any of these are present, automated governance alone is insufficient.

- Irreversibility: the AI decision cannot be walked back. Benefit termination, enforcement action, financial commitment.
- Consequence transfer: someone other than the decision-maker bears the outcome. Asymmetry creates a real accountability obligation.
- Distributional novelty: the case falls outside the training distribution. A human reviewer is not slower, they are more accurate.
- Value conflict: the decision requires a judgment the system has no reliable basis to make.
- Legal or regulatory accountability: the decision carries formal legal exposure that cannot be delegated to a model.

The inverse rule is operationally significant: where none of the five criteria apply, automated governance is the correct governance model. Real-time controls, guardrails, rollback capability, and automated monitoring are faster, more consistent, and more reliable than human reviewers who lack sufficient context. Applying committee review to low-consequence deployments wastes governance capacity and makes oversight of high-consequence deployments harder to sustain.

The Three Governance Clocks

The consequence-tier model identifies the type of oversight necessary for appropriate governance. The Three Governance Clocks address whether that oversight is functioning in real time. Most agencies cannot answer all three clock questions with specificity for a live agentic deployment.

Detection Clock: when something goes wrong in an agentic deployment, how long before the agency knows? In deterministic systems, detection is systematic. In agentic systems, detection depends on whether monitoring was designed into the deployment architecture or retrofitted after an incident. The most common finding from Stream Two intake: the detection mechanism is not visible to the people responsible for operating the system.

Intervention Clock: once the agency detects a problem, how long before a qualified human can halt or modify the deployment? Intervention authority that exists on paper but has never been exercised in a drill or a real incident is not functioning oversight. In multi-agent environments, where agency-managed agents interact with third-party SaaS agents, intervention becomes a cross-vendor coordination problem. Agencies that have not mapped the combinatorics of their agent interactions do not know where their intervention authority ends.

Vendor architecture dependency is a related Intervention Clock constraint that procurement governance has not yet systematically addressed. An agency that cannot modify or suspend an agentic deployment without vendor cooperation does not hold real **HALT** authority. The vendor holds it. Contracts that do not specify the agency's right to suspend, modify, or migrate a system without vendor consent are contracts that structurally impair the Intervention Clock. This is a contract term, addressable before award. Agencies accelerating AI procurement under time pressure are most at risk of omitting it.

Accountability Clock: after an autonomous action causes harm or significant error, how long before the responsible party is identified? Deterministic systems have clear accountability: someone approved the rule set. Agentic systems dissolve accountability across multiple approved layers. Agencies often discover they approved a process, not a decision, and no one owns the outcome. A useful test for any agentic deployment: can you identify a named individual who is accountable for an output the system produces without direct human review? If the answer requires consulting an organizational chart, the accountability clock is running.

PRACTITIONER RESEARCH: THREE FINDINGS

The litigation clock question: A practitioner at the June session raised whether litigation exposure should be treated as a fourth governance clock. When an AI action causes harm, how long before legal exposure crystallizes and what is the agency's window to act?

The multi-agent combinatorics question: Participants raised a governance gap that Stream Two research is documenting separately: in environments where agency agents, third-party SaaS agents, and vendor models interact, governing the full interaction space requires a different architecture than governing individual deployments. PII leakage across agent boundaries is the most immediate exposure; unauthorized scope expansion across agent interactions is the structural risk.

The accountability question: Accountability is not meaningful without real consequences. This maps directly to Stream Two's authority gap finding. Naming a reviewer is not the same as giving a reviewer the authority to act. Naming an accountable party in a governance document is not the same as ensuring that party has the operational access, the information, and the decision rights to be accountable in practice.

A Bifurcated Policy Environment

A presidential national security memorandum issued this week directs accelerated AI adoption across the Department of Defense (DoD) and the intelligence community (IC), with updated governance and

oversight requirements specific to national security applications. The directive creates a policy bifurcation that agency technology leaders need to address explicitly in governance documentation.

Some agencies operate across both national security and civilian contexts (including DoD components with citizen-facing functions, and contractors working in both environments.) For these organizations, the governing framework question is now a deployment-level decision, not an agency-level one. A deployment serving a civilian function remains under M-25-21 and the NIST AI RMF. A deployment serving a national security function operates under a separate accelerated timeline with different oversight requirements. Governance programs that have not distinguished between these two operating contexts at the deployment level are now producing documentation that does not accurately describe what governs each system. This has become a critical issue in search of resolution.

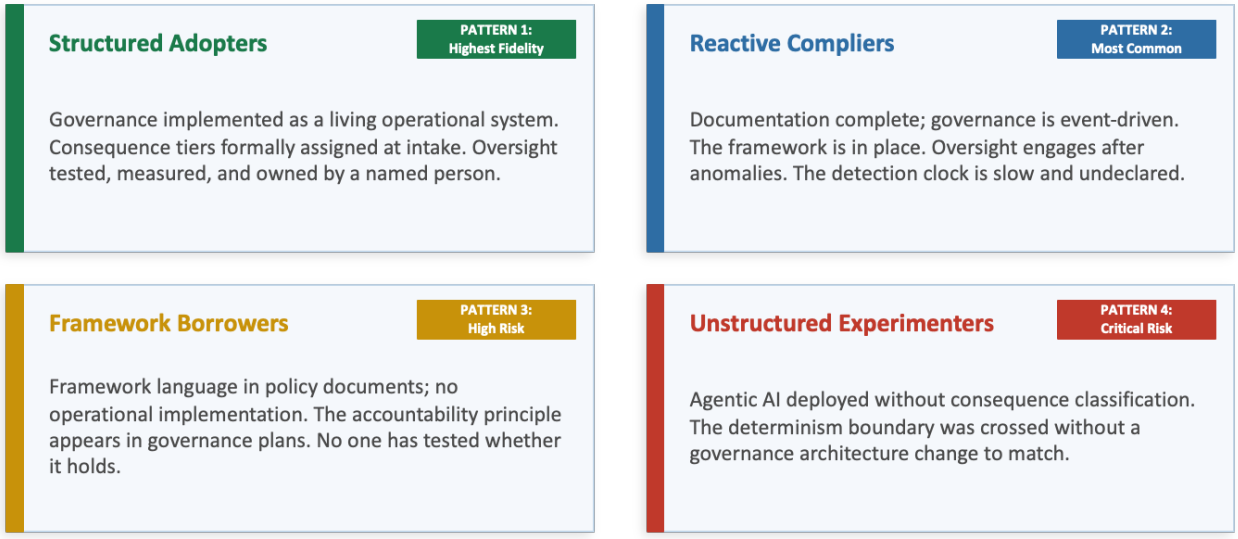
The Stream One implementation fidelity question applies directly: will agencies whose governance documentation cites the NIST AI RMF update that documentation to reflect which deployments operate under a different governing framework? Or will documentation remain static while the operational environment changes beneath it? What we have seen in the past with FISMA is an instructive precedent. Policy authority arrived before verification infrastructure in every prior governance cycle. Agencies that align their governance records to the actual governing framework for each deployment now are in a substantially stronger audit position when the verification cycle catches up.

STATE

The Four-Pattern Typology

Stream One research, examining NIST AI Risk Management Framework (RMF) implementation fidelity across federal and state agencies, has produced a four-pattern typology that describes what separates agencies whose governance holds under agentic conditions from those whose governance only looks good in documentation.

Structured adopters, the highest fidelity pattern, treat governance as a living operational system. Consequence tiers are formally assigned at intake for every new deployment. Oversight is tested, measured, and owned by a named person. Detection and intervention clocks have documented baselines. Stream One research finds very few agencies operating at this pattern for their agentic deployments. It requires a deliberate investment in operationalizing the framework (and not just documenting compliance with it.)



Reactive compliers, the most common pattern, have documentation in place and frameworks adopted. Governance engages after anomalies surface. The detection clock is slow because it depends on complaints or audits rather than automated monitoring. This pattern produces governance that is present but latent: the framework exists, the operational architecture to execute it does not.

Framework borrowers are at high risk. Governance language appears in policy documents. The accountability principle is stated in governance plans. No one has tested whether it holds in practice. When an agentic deployment causes harm, the documentation creates legal exposure without providing actual protection.

Unstructured experimenters present critical risk. Agentic AI is deployed without consequence classification. The determinism boundary was crossed without updating the governance architecture. A governance pattern that was adequate for earlier deterministic systems fails when applied to agentic deployments. Stream One research finds this is the early warning that most agencies do not recognize until something goes wrong.

The CAO Authority Gap in Agentic Contexts

May 2026 GIAG research posts established the authority gap at the designation level: most Chief AI Officer (CAO) designations carry compliance authority, not decision authority. In agentic deployments, the authority gap takes on an additional dimension. The same weakness the FITARA precedent demonstrated for Chief Information Officer (CIO) authority over IT acquisitions is now visible in CAO authority over agentic deployments.

The NASCIO model defines what functional authority over the transition from pilot to production looks like: an explicit right to authorize movement between exploration and production. Most state CAO designations do not carry an equivalent. CAO positions that have never delayed or modified a deployment decision may hold governance responsibility without governance authority. The compliance record does not distinguish between the two. The post-incident review will.

STREAM ONE: EARLY PATTERN FINDINGS

Four patterns are emerging consistently across Stream One structured interviews. These are directional findings from practitioner intake conversations and the public deployment record, not conclusions.

- Governance is front-loaded, not operational. Risk assessment and authorization receive substantial pre-deployment attention. Post-deployment operational monitoring is thin and frequently unassigned.
- The CAO role carries compliance authority, not decision authority. Most CAOs are responsible for documentation production. What they typically do not hold is authority to delay or modify a deployment the program office wants to proceed.
- Calendar triggers dominate behavioral triggers. Governance reviews occur when the compliance calendar requires them. They rarely occur when production system behavior generates a signal that warrants review.
- Human-in-the-loop language remains operationally undefined. The term appears in virtually every agency governance document reviewed. It almost never specifies which decisions require human review, who conducts it, what information the reviewer has, or what the documentation standard is.

LOCAL

Downstream Governance and the Agentic Threshold

Local government agencies are where the governance capacity gap is most acute and where informal AI adoption most consistently outruns formal governance. They are also the agencies most likely to receive AI systems through state or federal procurement channels without the governance architecture those systems require.

The HHS Administration for Children and Families is offering $6 million to state and local governments for predictive analytics in child welfare systems. This is federal funding accelerating algorithmic decision-making in high-stakes services for vulnerable populations. Consequence criteria one through four from the Stream Two intervention framework apply directly: irreversibility, consequence transfer to the affected population, distributional novelty in populations not well-represented in training data, and value conflicts inherent in child welfare decisions. A grant that funds deployment without specifying the oversight architecture those criteria require is transferring risk to the local agencies that receive it.

Local agencies that are downstream from state governance frameworks inherit the documentation layer of those frameworks. When state frameworks include operational architecture, funded oversight roles, current inventories, and defined reviewer authority, local agencies inherit a functional floor. When state frameworks consist primarily of documentation, local agencies inherit the documentation. The operational protection the documentation describes does not transfer with it.

For local agencies standing up agentic capabilities in child welfare, benefits routing, or permit processing: the NIST AI RMF is the most practical available baseline, used as operational architecture rather than compliance checklist. The value is in treating the five intervention criteria as deployment intake requirements, assigning a named individual to each clock, and defining what that individual can do when the clock is running.

PRACTITIONER SIGNALS: THREE TESTS FOR AGENTIC OVERSIGHT READINESS

These diagnostic questions emerged from a recent Stream Two practitioner exchange. They are presented as practical tools for any governance or oversight role in an agency with active agentic deployments.

Test One: Can You Run the Three Clocks on a Live Deployment?

Take one agentic deployment currently in production. Answer three questions: When this system takes an action outside expected parameters, how does the agency know, and how quickly? Who holds the authority to halt this deployment, and has that authority been exercised in a drill or a real incident? If this system produces an output that causes harm, who is the named accountable party? Three strong answers indicates oversight architecture that is functioning. Any partial answer identifies a gap. A “don’t know” answer is itself a governance finding: the oversight architecture is not visible to the people responsible for it.

Test Two: Has Governance Produced a Constraint?

Ask whether any governance review of a deployed agentic system has ever resulted in a concrete operational action: a scope restriction, a deployment pause, a vendor requirement, or a contract modification. A governance process that has never produced a constraint is generating records rather

than governance outcomes. This is the authority gap Stream Two research is documenting consistently across intake interviews. The absence of constraints is not evidence of governance maturity; it is evidence that the review process has not been positioned to produce them.

Test Three: Does Your Oversight Architecture Account for Scope Drift?

Ask whether the governance structure around a deployed system has been updated since initial deployment to reflect how the system is currently operating. A system deployed 18 months ago under governance built for its original scope requires updated governance if it has since added data integrations, expanded user populations, or taken on additional decision functions. Most governance structures have not been updated. A governance structure that does not track the system it governs is a documentation artifact. In agentic deployments, where scope can expand through model updates, new integrations, and task delegation without triggering a formal re-authorization, the drift between authorized scope and operational scope is frequently the widest governance gap.

APPLIED RESEARCH: STREAM TWO INTAKE AND INSTRUMENTATION

Stream Two practitioner intake is ongoing. The questions emerging most consistently from structured interviews align with the three tests above: where in the decision workflow does the human reviewer sit, what information does the reviewer have at the point of review, and what is the reviewer authorized to do with a finding.

The GIAG instrumentation suite, available for use at www.thinkcapital.org/tools.html provides four assessment instruments designed to operationalize these questions. The Agentic AI Governance Assessment evaluates a deployment description against the five intervention criteria. The AI Use Case Risk Tiering Wizard provides a five-dimension risk classification with a mandatory override floor for physical harm and civil rights cases. The Human Oversight Quality Index scores oversight mechanisms for a deployment in active production. The AI RMF Implementation Fidelity Checker runs a 20-question assessment across all four NIST functions. All tools are free to use with no registration required.

Practitioners working in agency governance roles with responsibility for agentic deployments are the participants Stream Two research needs most. If the three clocks framework or the five intervention criteria connect to what you are managing in your environment, reach out at research@thinkcapital.org or participate at thinkcapital.org/research.html.

FIVE QUESTIONS FOR PRACTITIONERS

These questions are designed for use in an oversight review meeting or governance assessment. They locate where the gap between agentic AI governance documentation and operational governance practice is most likely to be found.

1. For each agentic deployment currently in production: is there a named individual who holds **HALT** authority, and has that authority been tested? If the answer requires consulting documentation, the authority does not function as a real-time control.

2. Has the operating scope of any current agentic deployment changed since initial authorization? If so, was a governance review conducted that assessed whether the original consequence tier assignment still applies?

3. In multi-agent environments where your agency's systems interact with third-party vendor agents or SaaS platform agents: has your agency mapped the interaction boundaries? What is the governance mechanism for detecting PII leakage or scope expansion across agent boundaries?

4. When a governance review of a deployed system produces a finding, what is the defined escalation path? Who has the authority to require a vendor to modify or suspend system operation based on a governance finding? Has that authority ever been exercised?

5. For each agentic deployment procured from a vendor: does your contract specify the agency's right to suspend, modify, or require architecture changes independently of vendor consent? If not, your **HALT** authority is a policy commitment the vendor does not have to honor.

6. For your highest-consequence agentic deployment: can you identify the specific points in the process chain where a human decision is required before the system continues execution? Or does human review occur at the output boundary, after the consequential decisions have already been made?

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

Archive and publications: thinkcapital.org/publications.html.

Michael Bragen, Principal, ThinkCapital LLC | michael.bragen@thinkcapital.org | thinkcapital.org | thinkcapital.substack.com