THE COMMIT BOUNDARY

Runtime Governance Architecture for AI Systems

Closing Federal AI Governance Policy Gaps: A Runtime Governance Framework for Consequential AI

Published on LinkedIn at: https://www.linkedin.com/pulse/closing-federal-ai-governance-policy-gaps-runtime-framework-willis-56mbe/

John M. Willis

August 9, 2026

A Special Edition of The Commit Boundary

Artificial intelligence policy in the United States is no longer starting from a blank page.

Federal agencies are already subject to substantial requirements for AI governance, risk management, testing, monitoring, human oversight, accountability, cybersecurity, procurement assurance, and independent review. National-security policy goes further, requiring AI systems to be reliable, robust, steerable, controllable, consistent with law and policy, and accountable to established chains of command. International standards establish organizational AI management-system requirements. And recent NSA, CISA, ASD, and allied guidance has moved the security discussion directly into runtime, recommending continuous authentication and centralized policy decision points for individual agent actions.

These developments matter because they change the policy question.

The United States does not primarily face an absence-of-AI-governance problem.

It faces an **execution-governance gap**.

As AI systems move beyond generating information and become capable of invoking tools, coordinating workflows, delegating tasks, modifying records, changing access, spending funds, initiating transactions, interacting with critical infrastructure, or otherwise producing real-world consequences, an additional question arises:

What determines whether the exact action an AI-enabled system is about to cause is authorized and admissible under the conditions that exist at that moment?

Existing policy increasingly requires the ingredients needed to answer that question.

What it does not yet provide is an integrated engineering and assurance framework for answering it consistently, enforcing the answer at the boundary where consequence becomes real, and independently verifying that the enforcement cannot be bypassed.

That is the gap Runtime Governance Engineering is intended to close.

The Federal Government Already Requires Significant AI Governance

A useful starting point is OMB Memorandum M-25-21, *Accelerating Federal Use of AI through Innovation, Governance, and Public Trust*.

M-25-21 is important because it explicitly connects federal AI governance to **decisions and actions with consequences**.

Under the memorandum, an agency use of AI may qualify as “high-impact AI” when AI output serves as a principal basis for a decision or action having a legal, material, binding, or significant effect on rights or safety. That determination can apply whether or not a human participates in the decision or action.

For high-impact AI, agencies must implement minimum risk-management practices. These include pre-deployment testing, AI impact assessments, continuing monitoring, independent review, human oversight and intervention, accountability, and appropriate fail-safes. Where a high-impact AI use cannot comply and its risk cannot adequately be mitigated, the agency must discontinue the AI functionality.

This is already a significant governance requirement.

It means federal policy is not concerned solely with whether a model is powerful, accurate, or secure. It is concerned with the consequences of the decisions and actions for which AI serves as a principal basis.

That is an important foundation for Runtime Governance.

But M-25-21 primarily establishes **risk-management obligations around an AI use case and system lifecycle**. It does not define an end-to-end technical architecture for determining whether every exact consequential transition remains admissible immediately before it becomes operationally real.

That distinction becomes increasingly important as AI systems act with greater autonomy.

Federal Procurement Policy Adds Independent Evaluability

OMB Memorandum M-25-22 extends the governance requirement into federal acquisition.

For high-impact AI, contracts must support compliance with M-25-21. Agencies must be able to monitor and evaluate the performance, risks, and effectiveness of AI systems and services.

More significantly, vendors must provide agencies the access and time necessary for independent evaluation. Where vendors perform the testing instead, results must, where practicable, be detailed enough to be independently verified or reproduced.

This establishes an important policy principle:

Claims about consequential AI should be independently testable.

That principle should extend beyond model performance.

If an organization claims that an AI-enabled system is “human controlled,” “authorized,” “policy governed,” “non-bypassable,” or capable of refusing impermissible actions, those claims should eventually be susceptible to independent verification as well.

The key question should not simply be whether the required control exists.

It should be whether the claimed governance property actually holds when someone tries to violate it.

Executive Policy Is Simultaneously Accelerating AI Adoption and Security

The current Executive Branch policy environment reinforces this direction.

Executive Order 14179, *Removing Barriers to American Leadership in Artificial Intelligence*, reset federal AI policy toward accelerated innovation and adoption and revoked the prior EO 14110 framework. M-25-21 subsequently implemented that policy for federal agency AI use while expressly retaining governance and public-trust requirements. (**The White House**)

This matters because Runtime Governance should not be framed as an argument against accelerated AI deployment.

It is potentially an enabling architecture for it.

The more confidently an organization can demonstrate that consequential actions remain subject to legitimate authority and enforceable constraints, the more autonomy it may be able to deploy without responding to every risk by disabling capability entirely.

The June 2026 Executive Order *Promoting Advanced Artificial Intelligence Innovation and Security*—EO 14409—adds another dimension. It directs federal action around cyber defense, AI-enabled defensive tools, vulnerability coordination, critical-infrastructure protection, and benchmarking of advanced cyber capabilities in frontier AI models. (**The White House**)

These measures strengthen model security, system security, vulnerability management, and cyber resilience.

They do not, however, answer a different operational question:

Even if the AI system is secure, is this exact proposed action authorized to occur now?

Security and authority overlap.

They are not equivalent.

National-Security Policy Makes the Authority Problem Even Clearer

The requirement becomes particularly visible in NSPM-11, *Artificial Intelligence in the National Security Enterprise*, issued in June 2026.

NSPM-11 requires AI technologies adopted by the national-security enterprise to be designed to be **reliable, robust, steerable, and controllable**, and to operate in accordance with applicable laws, government policies, and guidance.

It requires rigorous testing, evaluation, validation, and verification.

It maintains responsibility and accountability in commanders, directors, and agency heads.

And it expressly requires AI use to remain consistent with civil liberties and protections afforded by the Constitution and applicable laws and regulations. (**The White House**)

NSPM-11 also requires standardized national-security AI Test, Evaluation, Verification, and Validation methodologies, including **conformity verification** for high-security AI systems. (**The White House**)

These are demanding requirements.

But consider what implementing them means in an autonomous operational environment.

If an AI-enabled system proposes a consequential action, how does the system determine that the action remains within the current operational authority of the relevant commander or principal?

How does it determine that delegated authority remains valid?

How does it distinguish authority from mere technical access?

How does it ensure that constitutional, statutory, safety, security, or mission constraints remain operative at the point of execution?

How does it prevent a human operator from approving something outside that person's own authority?

And how does an independent evaluator verify that an action inconsistent with those constraints cannot nevertheless be made to occur?

These are not merely model-evaluation questions.

They are execution-governance questions.

ISO/IEC 42001 Establishes an Important Management-System Layer

The same architectural distinction is visible internationally.

ISO/IEC 42001:2023 specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System within an organization. It provides a management-system structure through which organizations can establish AI policies, objectives, responsibilities, processes, risk management, and continual improvement. (**ISO**)

For organizations adopting or seeking certification against ISO/IEC 42001, these are substantive management-system requirements.

ISO/IEC 42006:2025 adds an independent-assurance dimension by specifying requirements for bodies that audit and certify AI management systems under ISO/IEC 42001. (**ISO**)

This is an important development.

But ISO/IEC 42001 is intentionally a management-system standard, not a prescribed execution architecture.

It does not require every organization to implement one particular control plane, policy engine, identity architecture, transaction gateway, or runtime decision mechanism.

Nor does it define a general architecture requiring each consequential state transition to be evaluated against current authority, qualified evidence, authoritative state, and all applicable governing invariants immediately before commitment.

That is not a deficiency in ISO/IEC 42001.

It is a layer distinction.

ISO/IEC 42001 establishes how an organization manages AI responsibly.

Runtime Governance Engineering asks how the resulting governance remains operationally effective **when an AI-enabled system is about to cause something**.

NSA, CISA, ASD, and Allied Guidance Is Moving Directly Toward Runtime

One of the strongest signals of this transition comes from the 2026 joint *Careful Adoption of Agentic AI Services* guidance issued by Australia's ASD/ACSC together with NSA, CISA, and other allied cybersecurity authorities.

The guidance recognizes that autonomous agent actions create new risks and states that those risks require updated governance policies and **continuous runtime authentication with centralized policy decision points for each action**.

It recommends limiting agent privileges, using just-in-time credentials for high-impact or privileged actions, authenticating agents with fresh cryptographic proofs, using cryptographic signing and integrity checks, continuously verifying identity and authorization at runtime, maintaining human oversight, establishing fail-safe behavior, and applying hard constraints that agents cannot override. (Cyber.gov.au)

This is an important convergence.

The allied cybersecurity community has moved beyond the assumption that authenticating an agent once, provisioning a credential, and permitting access to a tool is sufficient.

Identity and authorization increasingly need to be evaluated **during execution and at the level of individual actions**.

But even this leaves another important distinction.

Authorization Is Necessary. Operational Admissibility Is Broader.

Suppose an AI agent has a valid identity.

Suppose its credentials are current.

Suppose it has permission to call a particular API.

Suppose the API itself is secure.

The exact proposed transaction can still be impermissible.

The amount may exceed the principal's delegated authority.

The target may have changed.

The purpose may fall outside the permitted scope.

Required evidence may be missing or stale.

The operational state may have changed after approval.

A separation-of-duties constraint may apply.

A legal or safety requirement may prohibit the action.

A human approval may have expired.

Or the approving human may never have possessed authority to waive the relevant constraint in the first place.

Traditional access control might ask:

May principal P perform operation O on resource R?

Runtime Governance must be able to ask a more complete question:

May this exact consequential transition, proposed by this actor acting for this principal, against this target, for this purpose, under this delegation, using this evidence, under the current authoritative state and all applicable governing constraints, become operationally real now?

The distinction is fundamental.

Capability asks whether the system can perform the action.

Runtime Governance asks whether the action may occur now.

Prior authorization is not necessarily present authority.

NIST Already Provides Many of the Pieces

NIST is not absent from this problem.

Quite the opposite.

The NIST AI Risk Management Framework provides a broad structure for managing AI risk.

SP 800-53 provides an extensive catalog of security and privacy controls.

The Control Overlays for Securing AI Systems—or COSAiS—initiative is applying SP 800-53 controls to AI-specific cybersecurity risks. (**NIST Computer Security Resource Center**)

NIST's AI Agent Standards Initiative explicitly addresses AI agents capable of autonomous action and seeks industry-led standards and open protocols supporting secure, interoperable agent operation. (**NIST**)

The National Cybersecurity Center of Excellence is separately exploring software and AI-agent identity and authorization, including standards-based approaches to identifying agents and authorizing their access and actions. (**NCCoE**)

These efforts are all relevant.

Many of the controls needed to implement Runtime Governance substantially exist.

Identity controls exist.

Authorization controls exist.

Audit controls exist.

Integrity controls exist.

Risk-management processes exist.

Monitoring controls exist.

Human-oversight requirements exist.

Zero Trust concepts exist.

Secure-by-Design principles exist.

The missing issue is their **composition into an end-to-end execution-governance architecture**.

The currently published NIST portfolio does not yet establish, as a single integrated and independently testable system property, the full chain from:

authoritative governance → machine-evaluable governance → exact proposed consequential transition → current authority and delegation → qualified evidence and authoritative state → runtime admissibility determination → governed human escalation → decision-to-action binding → non-bypassable enforcement → durable evidence and independent verification.

That is a scope distinction, not a criticism of NIST's existing work.

The research underlying this article reaches the same conclusion:

The controls substantially exist. The requirement increasingly exists. The integrated architecture and assurance discipline do not.

Existing Law Also Provides Authority Without Defining the Complete Architecture

Federal law likewise does not begin from zero.

Congress has already assigned NIST substantial AI standards responsibilities.

15 U.S.C. § 278h-1 establishes NIST responsibilities for artificial-intelligence standards, frameworks, best practices, and related work. The broader National Artificial Intelligence Initiative statutory structure provides additional federal responsibilities around trustworthy AI, research, standards, and government use. (U.S. Code)

Existing laws governing privacy, civil rights, cybersecurity, financial transactions, healthcare, critical infrastructure, procurement, administrative authority, contracts, and sector-specific activities also continue to apply when AI is involved according to their respective scopes.

The problem is therefore **not that AI-enabled actions somehow exist outside the law**.

The problem is that law and regulation generally establish obligations, rights, prohibitions, authorities, procedures, and accountability without prescribing a common technical architecture for evaluating all of those constraints against an exact autonomous action at the instant of execution.

No current public federal framework identified in the underlying research expressly defines the national standards problem as the engineering of legitimate governance into the execution path of consequential autonomous action—from authoritative source,

through runtime determination and non-bypassable enforcement, to independently verifiable evidence.

That is the policy gap Congress can now address.

Governance Must Become Executable

Closing the gap requires more than placing a policy engine in front of an API.

Human governance is not automatically machine-executable.

Constitutions, statutes, regulations, agency directives, contracts, policies, delegations, safety requirements, and organizational rules are written for human institutions.

They contain jurisdiction.

Scope.

Definitions.

Exceptions.

Precedence.

Conflicts.

Temporal conditions.

Evidentiary standards.

Delegated authority.

Interpretive decisions.

A Runtime Governance system therefore requires a controlled process by which legitimate human governance becomes validated, approved, versioned, traceable, machine-evaluable governance.

This process can be understood as **Governance Compilation**.

Governance Compilation should not mean delegating legal interpretation to a language model.

Authorized legal, regulatory, policy, security, safety, and domain authorities remain responsible for interpretation and approval.

The engineering objective is to preserve that approved meaning through the transformation into runtime controls.

Ideally, an independent auditor should be able to trace backward from:

runtime decision → governing invariant → compiled requirement → approved interpretation → authoritative source.

If machines are going to enforce human governance, the process by which human governance becomes executable must itself be governed.

“Human in the Loop” Is Not an Architecture

The same issue applies to human oversight.

OMB requires suitable human oversight and intervention for high-impact federal AI.

NSPM-11 preserves responsibility in commanders, directors, and agency heads.

Allied agentic-AI guidance emphasizes continued human oversight.

But simply inserting a human approval button does not establish legitimate authority.

A meaningful system must still determine who the human is, what authority that person possesses, whether the authority is current, whether it was delegated, what scope and purpose it covers, whether separation-of-duties constraints apply, which conditions are non-waivable, which exact proposal is being approved, and whether the relevant state has changed before execution.

A human should therefore be treated as a **governed escalation authority**, not as an unrestricted override.

If Runtime Governance returns ESCALATE, the system should route the proposal to a qualified human authority, capture the resulting evidence, and then re-evaluate the proposed transition before commitment when material conditions could have changed.

This does not diminish human authority.

It preserves it.

A human click is not authority.

Legitimate, current, bounded authority is authority.

A Governance Decision Is Meaningless If It Can Be Bypassed

This leads to perhaps the most important architectural requirement.

A Runtime Governance mechanism must control the boundary where consequence becomes real.

A policy engine can produce the correct answer and still fail as governance if another execution path can bypass it.

A REFUSE decision means little if another interface can perform the same mutation.

An approval means little if it can be replayed against another target.

An authorization means little if action parameters can be changed afterward.

A runtime check means little if the system validates a proposed API request but not the state transition ultimately committed to the authoritative system.

The correct assurance question is therefore not merely:

Was the policy engine called?

It is:

Can the consequence occur without a valid governance determination?

For a qualifying Runtime Governance implementation, the answer should be no.

Every relevant consequence-producing path must be mediated.

A valid decision must remain bound to the exact proposal, target, tenant, conditions, and state for which it was issued.

And where material conditions can change between authorization and commitment, the system must be capable of re-evaluating admissibility at the point of commitment.

This is **non-bypassable consequence-boundary enforcement**.

Governance Claims Must Become Falsifiable

The federal policy environment is already moving toward independent evaluation and conformity verification.

Runtime Governance should build that requirement into the architecture from the beginning.

Suppose an organization claims:

“All high-impact AI actions require human approval.”

An assessor should not merely verify the existence of an approval workflow.

The assessor should attempt to defeat it.

Use a stale approval.

Change the target.

Modify action parameters.

Use an approver whose delegated authority has expired.

Create a race condition between approval and commitment.

Attempt an administrative bypass.

Attempt an emergency-mode bypass.

Use an alternative execution interface.

Substitute governance artifacts.

Supply stale authoritative state.

Replay evidence.

Try to produce the consequence after the system has issued REFUSE.

The most meaningful question in Runtime Governance assurance is a negative one:

Can an action that should not occur nevertheless be made to occur?

Independent verification should test whether inadmissible consequences are structurally prevented—not merely whether the required documentation, control descriptions, or policy statements exist.

Congress Can Close the Gap Without Creating Another AI Regulator

The policy solution does not require a new general-purpose federal AI regulatory agency.

It does not require Congress to prescribe one proprietary architecture.

And it does not require replacing the AI governance structures already being developed through OMB, NIST, CISA, the national-security enterprise, sector regulators, ISO, or industry.

Congress can instead assign a focused engineering and standards mission to institutions that already possess the necessary responsibilities.

Congress should direct the Department of Commerce, acting through NIST and the Center for AI Standards and Innovation, in coordination with the Department of Homeland Security through CISA and other appropriate agencies, to develop a **voluntary, technology-neutral national framework for Runtime Governance Engineering of consequential AI-enabled systems**.

The federal standards process should define the common properties that qualifying systems need to demonstrate while leaving organizations free to choose how they implement those properties.

Those properties should include authoritative governance and provenance; controlled Governance Compilation; machine-evaluable legal, constitutional, safety, security, and system invariants where appropriate; current authority and delegation; qualified runtime evidence; authoritative operational state; action-specific admissibility; bounded human escalation; decision-to-action and execution-target binding; commitment-time re-evaluation; structural refusal and safe degraded operation; non-bypassable enforcement; continuation and composite-action integrity; durable governance evidence; protection of the governance mechanisms themselves; runtime measurement; and independent conformity assessment.

The distinction is critical:

Standardize required properties, not products.

A federal framework might establish the normative requirement that every governed consequential transition receive a valid Runtime Governance determination before commitment.

It should not dictate whether an organization satisfies that requirement with a centralized control plane, distributed architecture, transaction gateway, service mesh, cryptographic capability mechanism, trusted hardware, embedded enforcement, or another implementation.

The requirement should be common.

The implementation should remain open.

That is **conformance standardization rather than implementation standardization**.

Start With Standards, Pilots, and Independent Testing

The initial framework should be voluntary.

NIST and CISA should then establish pilots involving appropriate high-impact federal AI and critical-infrastructure use cases.

Those pilots should focus on operational evidence.

Can authorization be reused after revocation?

Can a target be substituted after approval?

Can state change between authorization and execution without triggering re-evaluation?

Can a human exercise authority beyond the scope actually delegated?

Can another agent or interface bypass Runtime Governance?

Can the governance service fail open?

Can a composite workflow produce an impermissible result even if its individual steps appear permissible?

Can an independent assessor reconstruct exactly why the system permitted, refused, held, or escalated a particular transition?

And most importantly:

Can a consequential action that governing authority says must not occur nevertheless be made to occur?

Congress can then require a report addressing the framework, pilot results, unresolved standards gaps, conformity-assessment options, international standards alignment, and any additional statutory authorities shown to be necessary.

That sequencing is important.

Standards first.

Engineering evidence second.

Additional regulation where demonstrated gaps actually require it.

The underlying policy proposal recommends exactly this type of NIST/CAISI and CISA framework, including voluntary technology-neutral standards, reference architectures, test methods, conformity criteria, federal and critical-infrastructure pilots, and a subsequent report to Congress.

Closing Federal AI Governance Policy Gaps

The United States already has substantial AI governance.

OMB requires risk management, testing, monitoring, human oversight, accountability, independent review, and fail-safe mechanisms for high-impact federal AI.

Federal acquisition policy increasingly requires independent evaluability.

Executive policy is accelerating AI adoption while simultaneously strengthening cybersecurity, critical-infrastructure protection, and frontier-model assurance.

NSPM-11 requires national-security AI to be reliable, robust, steerable, controllable, accountable, and consistent with law, policy, constitutional protections, and established operational authority.

ISO/IEC 42001 establishes organizational AI management-system requirements, while ISO/IEC 42006 strengthens the associated certification and assurance ecosystem.

NSA, CISA, ASD, and allied authorities are already moving authentication and authorization into runtime and toward individual agent actions.

NIST is advancing AI risk management, security controls, AI cybersecurity overlays, autonomous-agent standards, identity, and authorization.

Existing law continues to govern the underlying rights, obligations, authorities, prohibitions, and consequences associated with AI-enabled activity.

These foundations should remain.

But increasingly autonomous systems create an additional engineering problem.

A model may propose an action.

The agent may have a legitimate identity.

The credential may be valid.

The tool may be approved.

The API may be secure.

The use case may have passed an impact assessment.

A human may previously have approved the workflow.

Every individual component may appear compliant.

And the exact proposed action may still be inadmissible now.

The remaining national-policy question is therefore increasingly precise:

May this exact AI-enabled action become operationally real under the authority, evidence, state, and governing conditions that exist at the moment of commitment?

A national Runtime Governance Engineering framework would establish how legitimate governing authority becomes executable; how current authority, delegation, evidence, state, and invariants are evaluated; how humans exercise governed and bounded authority; how decisions remain bound to the actions they govern; how consequence-producing paths are controlled non-bypassably; and how independent assessors determine whether those properties actually hold.

This does not require replacing the governance architecture already taking shape.

It requires connecting that architecture to execution.

Models propose.

Governance compilation makes legitimate authority executable.

Runtime Governance determines admissibility.

Humans exercise governed, bounded authority.

Enforcement controls consequence.

Independent verification proves the controls work.

The objective is not less capable AI.

It is **more capable AI operating under stronger, demonstrable control of consequence**.

Congress now has an opportunity to close that gap deliberately—before consequential autonomous execution becomes commonplace.