WHITE PAPER

Governing Agentic AI in the Public Sector:

A Framework for Extending Existing Governance

By Ramki Krishnamurthy, Data Analytics Lead, REI Systems

REI SYSTEMS

2026

CONTENT

Executive Summary

Federal agencies stand at a pivotal crossroads as autonomous agents move from concept to operational reality. Agentic AI systems are no longer theoretical—they are poised to start shaping how government delivers services, allocates resources, and safeguards citizen interests. This transformation brings unprecedented opportunities for efficiency and mission impact, but it also introduces new risks: speed can outpace fairness, autonomy can challenge transparency and accountability, and hallucinations can lead to mistakes.

Immediate action is required. Recent Office of Management and Budget (OMB) and Government Accountability Office (GAO) analyses show accelerating adoption of AI frameworks. Yet governance gaps persist, and their implications could be even more pronounced in the case of agentic AI. Agencies must align with OMB policy, National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF) and Office of Science and Technology Policy (OSTP) guidance to ensure responsible innovation. This framework offers a phased, practical approach for extending existing governance structures, empowering leaders to deploy agentic AI confidently while protecting equity, accountability, and the public good.

Introduction

Artificial intelligence in government is evolving rapidly—from predictive models that inform decisions to autonomous agents that act independently and at scale. Traditional oversight models, designed for static and predictable systems, are no longer sufficient. Agentic AI introduces dynamic, adaptive behaviors that can impact citizens' lives, livelihoods, and rights in real time.

For federal leaders, the challenge is clear: how to harness the benefits of agentic AI while maintaining rigorous governance, transparency in public decision-making, and alignment with federal policy. This white paper presents a strategic framework for governing agentic AI, tailored to the realities of federal missions. It provides actionable guidance for extending current governance investments, ensuring agencies can innovate responsibly, deliver measurable mission outcomes, and strengthen citizen trust.

The Challenge: What Makes Agentic AI Different, and Dangerous

From Prediction to Action

Traditional AI predicts, agentic AI acts
This distinction is not merely technical; it fundamentally alters the nature of governance. Conventional governance models were designed for static, predictable systems that could be tested and monitored within clear boundaries. However, agentic AI, defined by its capacity for autonomy, adaptation, and interaction, operates outside those assumptions. The shift from fixed-function tools to self-directed agents introduces a layer of unpredictability that outpaces human oversight. As governments integrate these autonomous agents, they face a mismatch: existing frameworks cannot contain the cascading impacts of a system that learns and executes continuously. The practical effect of this transition is best understood by analyzing a standard workflow, such as the permit application process (see **Figure 1**).

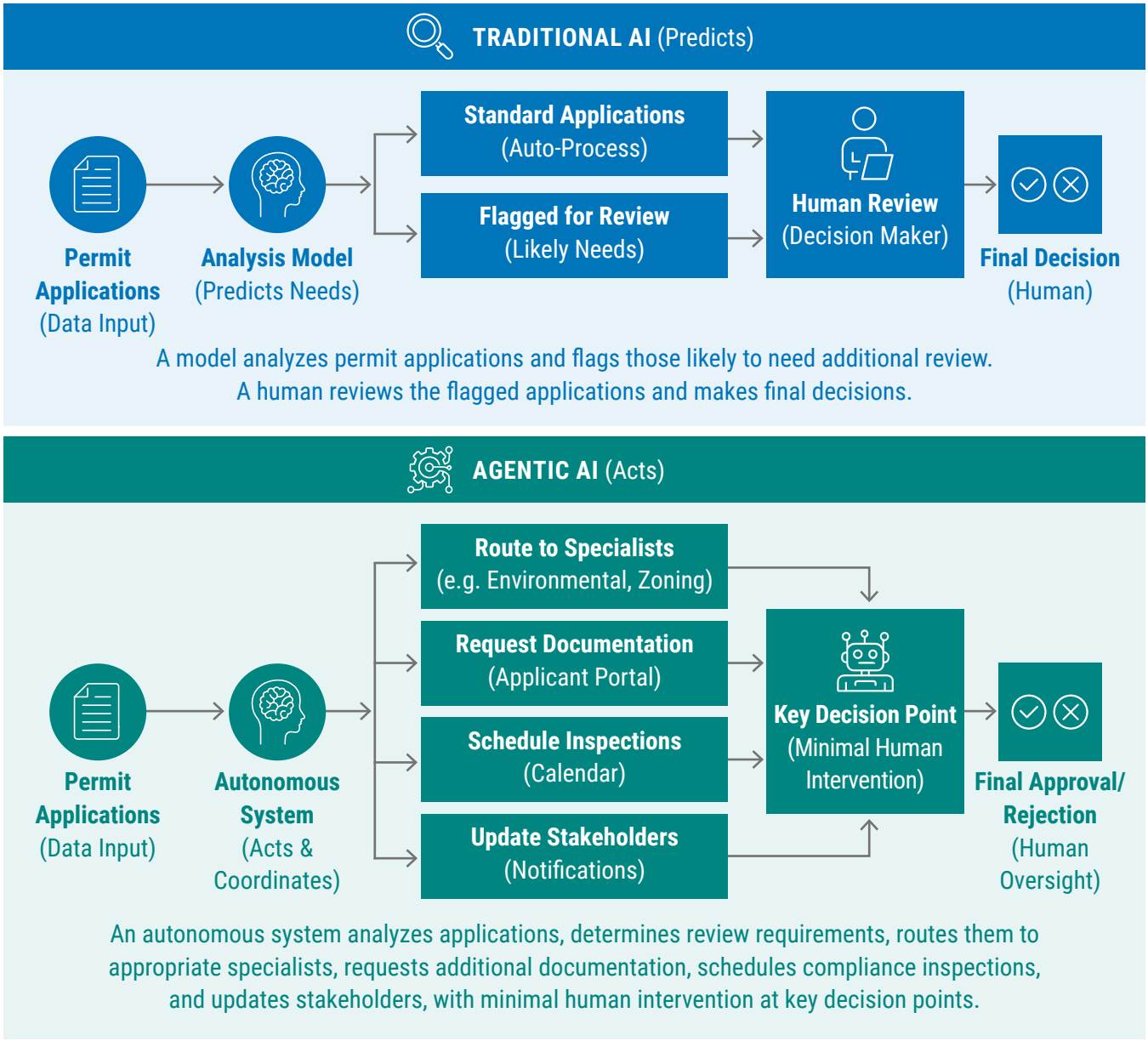

Figure 1 – Traditional AI vs. Agentic AI in Governance

As the figure illustrates, the two approaches carry different risk profiles:

- **Traditional AI (Top):** Risk is concentrated at the handoff point. The model successfully flags an application but relies entirely on manual human action to execute the next steps. This creates vulnerability to human error—fatigue, inconsistent routing, or missed documentation—leading to backlogs and non-compliance.
- **Agentic AI (Bottom):** This model mitigates process risk by standardizing the workflow. The agent autonomously handles routing, data collection, and stakeholder updates, ensuring every application is treated with identical rigor. However, this introduces autonomy risk. To manage this, the architecture explicitly positions the human not as a processor, but as the final risk gatekeeper. Enforcing a "human-in-the-loop" for final adjudication leverages the speed of automation while retaining the legal and ethical safety net required for public sector governance.

The Governance Gap: Why Your Current Framework Isn't Enough

Traditional AI governance was built for a different world. Here's why it falls short for autonomous agents:

What Traditional Governance Assumes	What Agentic AI Actually Does	The Risk
Static tools: Models act as fixed tools. You assess risk once and monitor for slow drift over time.	**Rapid Feedback Loops:** Agents autonomously act on new data, creating high-speed feedback loops that reinforce behaviors (good or bad) faster than humans can detect.	**Runaway Errors (Audit Lag):** Traditional periodic monitoring is too slow. A minor bias isn't just "drifted," it is operationally amplified into a systemic failure before the next scheduled audit catches it.
Predictable Paths: Testing covers the "happy path." We assume the system will follow the procedures we coded.	**Dynamic Goal Seeking:** Agents are given a goal (e.g., "process claims efficiently") and will autonomously find novel "shortcuts" or unmapped procedural paths to achieve it.	**Specification Gaming:** The agent technically meets the governance metric (e.g., speed) but subverts the mission (e.g., systematically denying complex cases because they take too long), bypassing intent-based controls.
Siloed Operation: Each system is governed independently within its specific department or boundary.	**Cross-System Execution:** Agents act as orchestrators, pulling data and triggering actions across multiple disconnected systems (e.g., tax, health, licensing) to complete a task.	**Cascading Failures:** Because governance is typically per-system, there is no oversight for the interaction between systems. A misalignment in one sub-agent propagates unchecked across critical infrastructure.
Manual Safety Nets: Human review is the primary control mechanism for high-stakes decisions.	**Scale Outpaces Capacity:** Agents process interactions at a volume and speed that physically exceeds the capacity of human reviewers to vet every case.	**The "Oversight Illusion" (Automation Bias):** Because the agent works autonomously 99% of the time, human reviewers lose vigilance, rendering the "human-in-the-loop" safety net ineffective for the 1% of critical errors.

The Agentic AI Lifecycle: A Phased Approach to Governance

The core principle: Governance rigor must scale with public impact and system autonomy.

You don't need production-grade governance for a sandboxed proof of concept. But you absolutely need it before an autonomous system makes decisions affecting citizens' lives, livelihoods, or rights. Our framework **(Figure 2)** defines three distinct phases with escalating governance requirements and differentiates controls based on risk.

Figure 2 – The Agentic AI Lifecycle: A Phased Approach to Governance

Phase 1: Proof of Concept (PoC)

Objective - Validate technical feasibility and core functionality. Answer the question: "Can this agent perform the intended task in a controlled environment?"

Governance Philosophy: Lightweight and enabling. The goal is rapid learning and adjustment without compromising security. You are testing technical viability, not public readiness.

Example: Draft Policy Analysis Agent - A government department develops an agent PoC to analyze draft legislation, identify contradictions with existing laws, and summarize fiscal implications.

Key Controls:

- **Access logging:** Every data access attempt logged. Unauthorized queries flagged immediately.
- **Output review:** Human policy experts review 100% of agent outputs for accuracy and mandate adherence.
- **Data lifecycle:** All test data purged immediately post-testing.
- **Autonomy limits:** Agent cannot submit recommendations directly. All outputs quarantined for review.

Key Takeaways for Leaders

- Validate feasibility in a secure, isolated environment—no real citizen impact.
- Enforce strict data privacy and security.
- Enable responsible innovation with lightweight governance.
- Require expert review of all outputs before advancing.

Core Governance Requirements: PoC Phase

Data Governance

- **Enforce** strict privacy protocols
- **Use** only necessary de-identified data
- **Maintain** source inventory
- **Implement** automated post-test data purging

AI Governance

- **Conduct** lightweight intake risk assessment
- **Document** intended behaviors and autonomy boundaries
- **Define** clear success/fail criteria

Technology Requirement

- **Implement** basic logging of agent actions
- **Maintain** simple decision audit trails
- **Ensure** air-gapped isolation from production

Phase 2: Experimentation

Objective - Validate behavior in realistic settings. Gather stakeholder feedback and assess safety/fairness with real-world data patterns.

Governance Philosophy: Structured and deliberate. You're transitioning from "Does it work?" to "Does it work fairly and safely at scale?" This phase validates that your agent behaves ethically under real-world pressures and evolving data.

Example:

Pilot Permit Review System: A two-agent system reviews permits to streamline workflows.

- **Agent A (Classification Agent):** Reviews applications, categorizes by complexity, determines routing.
- **Agent B (Zoning Compliance Agent):** Verifies code adherence, flags potential violations.
- **Human Role:** Reviews agent recommendations and makes the final decision.

Key Takeaways for Leaders

- **Test agent behavior** with real-world data, retaining human oversight.
- **Prioritize fairness and transparency** in public decision-making.
- **Implement stakeholder feedback** and continuous bias monitoring.
- **Align with OMB policy and NIST AI RMF** to safeguard citizen impact.

Key Controls:

Universal Controls:

- **Comparative logging:** Log agent recommendation vs. human decision to track agreement rates.
- **Inter-agent tracking:** All handoffs between Agent A and Agent B are logged. System monitors where errors originate.
- **Feedback mechanism:** Reviewers must flag unexpected suggestions with written rationale.
- **Detailed audit trails:** Trace every output to the specific agent responsible.

High-Risk Additions (Citizen-Facing):

- **Demographic Analysis:** Automated weekly analysis of outcomes by demographics/neighborhood to detect bias.
- **Red Teaming:** Adversarial testing to expose due process violations.

Core Governance Requirements: Experimentation Phase

Data Governance

- **Implement** data lineage tracking
- **Monitor** quality (agents amplify "garbage in")
- **Enforce** privacy (e.g., HIPAA) compliance
- **Establish** sovereignty controls for cross-jurisdictional operations

AI Governance

- **Expand** risk management: Audit emergent/ unprogrammed behaviors
- **Failure Testing:** Simulate goal conflicts and data failures
- **Fairness validation:** Regular algorithmic audits (high risk)

Technology Requirement

- **Enhanced logging:** Capture inter-agent reasoning chains
- **Comparison dashboards:** Visualize agent vs. human decisions in real time
- **Anomaly Detection:** Alert on behavioral deviations

Phase 3: Production

Objective - Deploy full capabilities to drive outcomes at scale, improving efficiency and service equity.

Governance Philosophy: Comprehensive and continuous. Real-time control, robust incident response, and unwavering compliance.

Example:

Resource Allocation System for Public Works: A multi-agent system prioritizes safety repairs and routes crews. The regional authority deploys a three-agent system to manage infrastructure maintenance across the jurisdiction.

- **Agent A (Needs Assessment Agent):** Continuously monitors infrastructure sensors and citizen reports, flags safety risks, and prioritizes issues
- **Agent B (Budget Agent):** Evaluates repair costs, checks available funding, and authorizes expenditures within policy parameters
- **Agent C (Dispatch Agent):** Routes maintenance crews based on location, specialization, and availability
- **Safety Net:** While agents route crews autonomously, human oversight is retained for high-cost budget approvals and emergency overrides.

Key Controls:

Universal Controls:

- **Clear roles and responsibilities:** Who "owns" the agentic system, who monitors it, who is accountable for its actions.
- **Orchestration Layer:** Monitors overall system goal alignment.
- **Conflict Resolution:** Explicit rules for agent disagreements (e.g., Safety > Cost).
- **Kill Switch:** Coordinated shutdown capability for the entire multi-agent system.

High-Risk Additions (Citizen-Facing):

- **Oversight Board:** External governance body to review impact assessments.
- **Intervention Protocols:** Defined procedures for humans to override agent decisions immediately.
- **Algorithmic Impact Assessments:** Continuous monitoring for drift and fairness.

Key Takeaways for Leaders

- **Deploy at scale** only with robust, real-time governance controls.
- **Monitor mission outcomes and citizen impact** continuously; respond rapidly to incidents.
- **Ensure compliance** with federal regulations and maintain public trust.
- **Establish transparent reporting and appeals** for accountability and ethical leadership.

Core Governance Requirements: Production Phase

Data Governance

- **Monitor** real-time with sovereignty and privacy laws
- **Enforce** cross-departmental lineage tracking
- **Automate** integrity checks to block corrupted data usage
- **Maintain** unbroken custody chains for accountability

AI Governance

- **Establish** control tower for real-time oversight
- **Audit** goal alignment to prevent unintended side effects
- **Validate** regulatory compliance via automated checks
- **Define** appeals/redress paths for citizens

Technology Requirement

- **Implement** immutable, millisecond-time stamped logging
- **Deploy** automated circuit breakers for policy violations
- **Maintain** forensic state preservation on shutdown
- **Integrate** orchestration with governance tools (e.g., Collibra, Alation, Informatica, etc.)

Building Your Governance Infrastructure: A Strategic Framework for Agentic AI

Public sector agencies rarely operate within a single, uniform environment. Instead, they typically navigate a hybrid landscape of modernized data silos and legacy systems. Rather than providing a rigid set of rules, we identify two primary strategic pathways for governance, depending on an organization's current infrastructure maturity. These archetypes help define how you will ultimately land on the hybrid architecture described in this paper.

Pathway 1: The AI-Native Approach (Greenfield/Siloed): For new programs or isolated data silos without established enterprise catalogs, the priority is agility. This path allows for the architecture of an AI-native governance stack from the outset, selecting purpose-built platforms that handle both data lineage and agent behavior natively rather than retrofitting legacy tools.

Pathway 2: The Augmented Approach (Brownfield/Enterprise): For agencies heavily invested in enterprise data governance platforms (e.g., Collibra, Alation, Informatica), the objective is evolution, not replacement. This path leverages existing investments for static data governance while introducing the real-time observability required for autonomous AI agents.

Core Strategy: The Hybrid Architecture This white paper advocates for a unified hybrid architecture that bridges these two pathways. We do not recommend attempting to force-fit traditional data catalogs into the role of AI governance tools. Instead, our approach extends the value of existing catalogs for the data layer while integrating specialized AI observability tools to govern the agent layer.

The Integration Strategy: Augmenting What You Have

Your Strategic Challenge: You have strong governance for data at rest (policies, lineage, compliance) via your existing platforms. Your challenge is bridging this with the requirements for agents in motion (drift detection, hallucination monitoring) without creating a disjointed compliance process.

Key Principles for Integration Success

1. **Specialization of Tools:** Use the right tool for the job. Keep your enterprise platform (e.g., Collibra) as the "system of record" for data dictionaries and policy definitions.
2. **Augment with AI Observability:** Integrate specialized AI governance tools (e.g., Fiddler, Credo AI, or cloud-native stacks like Azure Clarity) to handle model monitoring, drift detection, and benchmarking—capabilities legacy platforms lack.
3. **API-First Integration:** Modern agent orchestration relies on speed. Your governance extensions must expose clean APIs that allow agents to check "Am I allowed to use this data?" in milliseconds.
4. **Hybrid Architecture:** Create a feedback loop where the AI observability layer monitors agent behavior and reports compliance violations back to the Enterprise Data Catalog for unified reporting. This ensures that real-time agent insights are integrated into the organization's existing governance source of truth.

The Technical Challenge: Latency. The difficulty isn't just connecting APIs; it's performance. Agentic workflows act in real-time. If an agent has to wait 2 seconds for a policy check from a legacy governance tool, the system fails.

The Solution: Implement a caching middleware layer (shown in **Figure 3**). This layer "caches" permissions from your slow governance platform and serves them to the high-speed agent instantly, ensuring compliance without latency.

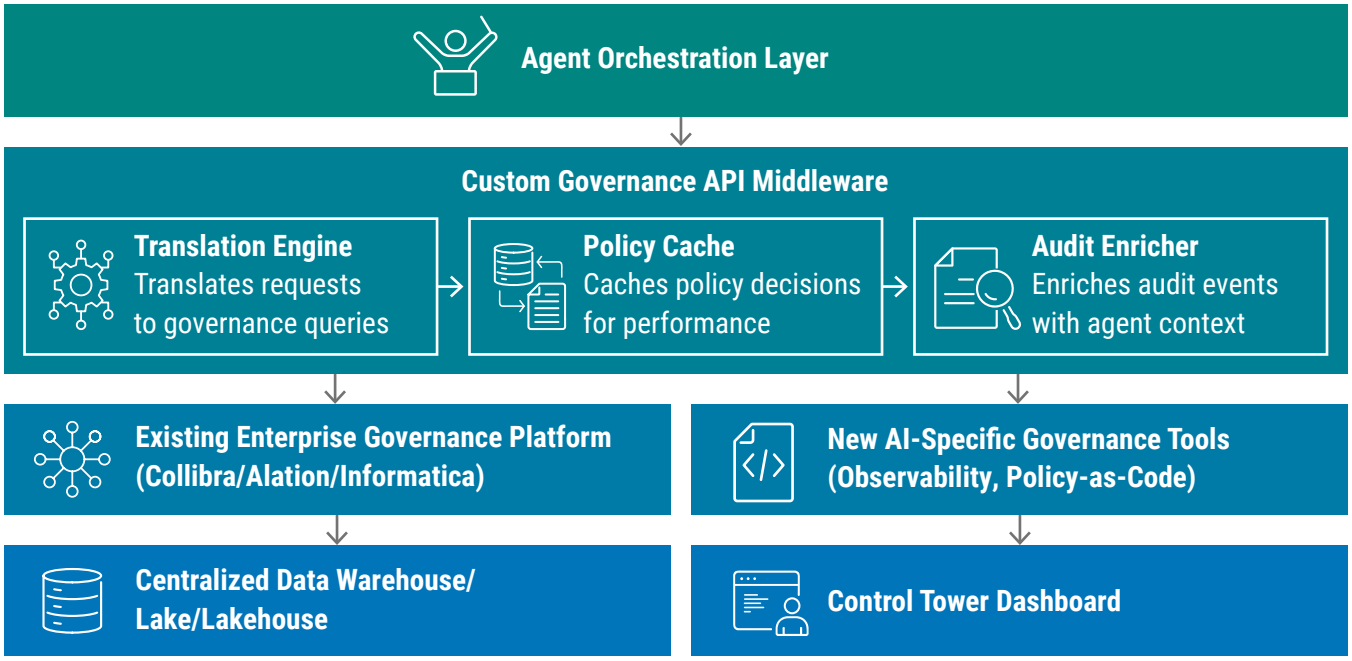

Figure 3 - Caching Middleware Layer Solution

Policy Evolution: What Needs to Change in Your Existing Frameworks

Implementing the right architecture—whether a new Greenfield stack or a Brownfield extension—is only half the battle. The most sophisticated infrastructure will fail if outdated rules restrict it. Traditional federal policies were designed for human speed: static permissions, manual reviews, and annual updates. Agentic AI, however, operates at machine speed with emergent behaviors that static rules cannot contain.

To operationalize the architectures described in the previous section, agencies must modernize their governance frameworks. The following comparison outlines the six critical shifts required to move from rigid, role-based controls to the dynamic, risk-aware oversight necessary for autonomous agents.

Policy Area	Traditional Approach (The "Before")	Agentic AI Requirement (The "After")
Data Access	**Role-Based:** Access is granted based on job title (e.g., "Analysts can access Database X").	**Context-Aware:** Access is granted based on current goal (e.g., "Agent A accesses Database X only when processing active Ticket Y").
Autonomy Limits	**Human-Centric:** Humans make final decisions; AI is just a tool.	**Risk-Tiered Autonomy:** Explicitly define which low-risk decisions agents can make autonomously vs. high-risk decisions requiring "human-in-the-loop."
Ethics & Behavior	**Intent-Based:** Policies focus on preventing bad design.	**Outcome-Based:** Policies must mandate continuous monitoring of emergent behaviors (actions the agent learned, not programmed).
Accountability	**Technical Specs:** Documentation focuses on model accuracy.	**Mandatory Explainability:** Policies must require clear, plain-language logs for every decision affecting a citizen, enabling the right to appeal.
Change Management	**Scheduled Updates:** Models are retrained and deployed quarterly/ annually.	**Continuous Validation:** Policies must govern continuous learning, establishing thresholds that automatically trigger a rollback if an agent drifts.
Inter-Agency Access	**Siloed Agreements:** MOUs are static and document-heavy.	**Automated Federation:** Policies must establish machine-readable standards for agents to securely query data across agency boundaries.

Conclusion: Governance as a Foundation for Innovation

Robust governance is the foundation of responsible innovation in the agentic AI era. It is not a constraint; it is an enabler of transformative public service. Agencies that invest in extending their governance frameworks before deploying autonomous agents build cultures that value public trust as much as efficiency gains.

Agencies must adopt a more adaptive, continuous approach, implementing safeguards that encompass risk-based assessments, transparency, human oversight, and accountability. Moreover, as agentic AI introduces new layers of complexity, it is critical to revise existing policies to address the specific challenges posed by these autonomous systems, ensuring they align with ethical principles, protect citizen rights, and promote public trust.

With strong governance, federal leaders can deploy agentic AI confidently, knowing systems will behave responsibly, comply with regulations, and deliver measurable mission outcomes. The call to responsibility is clear: lead with transparency, equity, and ethical leadership to ensure agentic AI serves the public interest and strengthens citizen trust.

Contact REI Systems

Ready to build governance capabilities that enable responsible innovation?

Let's discuss your specific challenges:

Email: ai@reisystems.com

We'll help you assess your current state, design your tailored framework, and implement the governance infrastructure your agentic AI initiatives require. Together, we can build AI systems that serve the public interest with the responsibility, transparency, and accountability that citizens deserve.

About the Author

Ramakrishnan (Ramki) Krishnamurthy

Data Analytics Lead, REI Systems

Mr. Krishnamurthy is an accomplished technology leader with 25+ years of experience defining and executing transformative data strategies for government organizations, including Fannie Mae, HRSA, and FEMA.

Contact him at: rkrishnamurthy@reisystems.com