Aquia

Federal Artificial Intelligence Guardrail Framework

Last updated: March 31, 2026

AI adoption across government is accelerating faster than the frameworks needed to govern it.

There is currently no unified, government-wide AI adoption framework. Agencies are on their own to navigate a threat landscape that includes algorithmic bias, citizen data exposure, adversarial AI attacks, and procurement failures while simultaneously under pressure to deliver mission results. The result is fragmented deployments, compliance gaps, and legal exposure, including violations of the Privacy Act of 1974.

The “Federal AI Guardrail Framework” provides security practitioners at federal agencies with a practical, actionable playbook for responsibly adopting AI, grounded in established standards.

www.aquia.us | federal@aquia.us

Table of Contents

Executive Summary

The Federal AI Guardrail Framework is a practical resource for information security practitioners, compliance leads, and agency technology officers navigating the responsible adoption of artificial intelligence in federal environments. It provides actionable guidance, risk mitigation strategies, and compliance mappings to help agencies integrate AI safely, ethically, and in accordance with established standards — including ISO/IEC 42001:2023, ISO 27001:2022, and the NIST AI Risk Management Framework (RMF) 1.0.

Federal agencies are deploying AI at an accelerating pace, yet the absence of a unified, government-wide framework has left many without the skills, controls, or governance structures needed to manage the associated risks. This document addresses that gap by offering a structured approach that enables responsible AI scaling while ensuring compliance with established federal standards.

The framework begins with an **AI Implementation Scoping** model that helps agencies classify their AI adoption into one of three scenarios: **Buying AI** (off-the-shelf SaaS and commercial API solutions), **Building with AI** (pre-trained models enhanced with Retrieval-Augmented Generation, or RAG), and **Owning AI** (custom model training and deployment). Each scenario carries a distinct risk profile, threat surface, and set of applicable controls, which are summarized in an **AI Scoping Matrix** to guide agencies before they design any security architecture.

The **Risk Mitigation** section identifies six critical federal risk areas and maps each to specific countermeasures:

1. The lack of a standardized government-wide framework, addressed through an internal NIST AI RMF Overlay and agency-wide acceptable use policy
2. Algorithmic bias, transparency, and explainability, addressed through an inference audit trail using chain-of-thought prompting and regular bias pressure testing
3. Protection of citizen data privacy, addressed through zero trust data enclaves with local PII scrubbing and ephemeral context windows
4. Procurement and lifecycle management of non-compliant deployments, addressed through an AI bill of materials (AIBOM) requirement and kill-switch protocol
5. The absence of a unified AI-as-a-Service platform, addressed through a sovereign AI shared service (SAISS) with a multi-tenant internal AI gateway
6. Scaling AI while maintaining public trust, addressed through a human-in-the-loop tiered response that classifies AI actions by impact level

A **Key Considerations for AI Adoption** section guides agencies through scoping, component inventory, resource assessment, and risk/trade-off analysis before implementation. The document then maps the framework's controls and guardrails to the three primary compliance standards: ISO/IEC 42001:2023 (AI management systems), ISO/IEC 27001:2022 (information security management), and NIST AI RMF 1.0 (covering the four discrete functions of map, measure, manage, and govern) — and provides concrete implementation actions for each.

Agencies that apply this framework will be equipped to evaluate their AI implementation scenario, select the appropriate guardrails, satisfy compliance requirements, and establish ongoing monitoring and governance.

Problem Statement

Government agencies are developing their AI capabilities, but progress is not keeping pace with the technology's evolution. The absence of appropriate training leaves agencies without the skills needed to responsibly adopt and manage AI and the risks that come with it.

Some of the top vulnerabilities associated with AI facing government agencies are:

1. **Federal agencies lack a shared, secure foundation for AI adoption.** An AI-as-a-Service (AIaaS) platform is needed to establish robust security and privacy guardrails, enable resource sharing, build durable and reusable development investments across agencies, and deliver beneficial AI-driven public services.
2. **Federal agencies face significant challenges in the secure and ethical adoption of AI** due to the lack of a standardized, government-wide framework for safety, security, and ethical governance.
3. **Current procurement and development processes are often ill-equipped to handle the unique lifecycle and safety requirements of AI systems**, leading to fragmented and potentially non-compliant deployments.
4. **The urgency to mature strengths to offset the AI-as-a-Service risks identified above is increasing faster than less nimble institutions are equipped to address.** Adoption requires a cohesive structure that enables rapid, yet responsible, scaling of AI capabilities while maintaining public trust and adhering to legal mandates.

AI Implementation Scoping

Before applying the right guardrails, agencies must first determine which type of AI implementation they are undertaking. The risk profiles, threat vectors, and applicable controls differ fundamentally depending on whether an agency is buying, building, or owning AI. Jumping directly into technical controls without this scoping step risks misallocating security resources and leaving critical threat vectors unaddressed.

AI adoption is frequently announced as a single initiative of "integrating AI" but this phrase conceals a wide range of architectural and procurement decisions, each with distinct security implications.

The three most common scenarios federal agencies encounter are described below:

1. Buying AI (consumer/SaaS applications)
2. Building with AI (pre-trained models and retrieval-augmented generation (RAG))
3. Owning AI (Custom Model Training)

Scenario 1
Buying AI (Consumer/SaaS Applications)

Many agencies begin their AI journey by adopting existing SaaS AI solutions. For example, using Microsoft Copilot, commercial API access to frontier models, or government-procured AI chat interfaces. While this approach offers the lowest barrier to entry, it introduces significant security and compliance considerations specific to the consumer/SaaS model.

Key Security Considerations:

- **Shared Responsibility Model:** The provider handles model security and inference infrastructure, but the agency remains responsible for prompt engineering guardrails, data classification policies, and use-case governance.
- **Data Flow Control:** When employees use consumer AI tools, sensitive data may flow outside the agency perimeter. Data loss prevention (DLP) policies and clear acceptable use policies (AUPs) are essential.
- **Terms of Service and Legal Review:** AI providers differ in how they handle and may use submitted data. A thorough legal and security review of vendor terms is required before organizational adoption.
- **Primary OWASP Threats*:** ASI03 (Identity and Privilege Abuse), ASI09 (Human-Agent Trust Exploitation).

*See page 8 for a list of 10 specific threats that must be mitigated for safe AI adoption, per the OWASP Top 10 for Agentic Applications for 2026.

Scenario 2

Building with AI (Pre-Trained Models + RAG)

Agencies seeking more control often adopt pre-trained models through cloud AI services (e.g., Amazon Bedrock, Azure OpenAI Service, Google Vertex AI) and enhance them with Retrieval-Augmented Generation (RAG) to incorporate proprietary or agency-specific data without fine-tuning the base model. This is a common pattern for knowledge bases, Q&A systems, and document summarization tools.

Key Security Considerations:

- **Vector Database Security:** RAG implementations store embeddings in vector databases that require proper access controls, encryption, and monitoring.
- **Prompt Injection Protection (ASI01, ASI06):** RAG pipelines are particularly vulnerable to prompt injection attacks that can extract information from the knowledge base or corrupt the retrieval context. Input sanitization and output filtering are essential in this scenario—more so than in consumer SaaS deployments where the agency does not control retrieval.
- **Authentication and Authorization:** The RAG system must enforce access controls on retrieved documents based on individual user permissions, especially when documents span multiple classification levels.
- **Primary OWASP Threats:** ASI01 (Agent Goal Hijack), ASI06 (Memory & Context Poisoning), ASI02 (Tool Misuse and Exploitation).

Scenario 3

Owning AI (Custom Model Training)

Some agencies, particularly those with unique mission data or stringent classification requirements, opt to train their own models or extensively fine-tune existing ones on controlled datasets. This scenario provides the highest level of data control but introduces the greatest infrastructure complexity.

Key Security Considerations:

- **Training Data Security:** The entire training corpus requires protection, including secure data pipelines, storage, and retention policies.
- **Model as a Security Asset:** The model itself must be protected from extraction, theft, or unauthorized access. It represents a significant investment and may encode sensitive information.
- **Supply Chain and Pipeline Security (ASI04, ASI05):** Custom training pipelines involve third-party libraries, frameworks, and compute infrastructure, each of which is a potential attack surface.
- **Primary OWASP Threats:** ASI04 (Supply Chain Vulnerabilities), ASI05 (Unexpected Code Execution), ASI10 (Rogue Agents).

AI Scoping Matrix

The table below summarizes the three scenarios, their risk profiles, and the guardrails most critical to each. Agencies should identify their scenario before proceeding to the "Risk Mitigation" section.

Dimension	Scenario 1: Buying AI (SaaS/Consumer)	Scenario 2: Building with AI (RAG + Pre-trained)	Scenario 3: Owning AI (Custom Training)
Examples	ChatGPT, Microsoft Copilot, Claude.ai	Amazon Bedrock, Azure OpenAI + RAG, Vertex AI	Agency-trained models, fine-tuned LLMs on classified data
Primary Risks	Data leakage, ToS compliance, PII in prompts	Prompt injection (ASI01, ASI06), vector DB access control	Training data security, model theft, pipeline vulnerabilities
Key Controls	DLP, AUP, ToS review, access governance	PII scrubber, vector DB access controls, output filtering	Data lineage (DVC), AIBOM, training pipeline security
Infrastructure Risk	**Low**	**Medium**	**High**
Data Leakage Risk	**High**	**Medium**	**Low**
Most Critical OWASP Threats	ASI03 (Identity Abuse), ASI09 (Trust Exploitation)	ASI01 (Goal Hijack), ASI06 (Memory Poisoning), ASI02 (Tool Misuse)	ASI04 (Supply Chain), ASI05 (RCE), ASI10 (Rogue Agents)

How to use this matrix: Identify your agency's scenario (or mix of scenarios) before proceeding to the Risk Mitigation section. The OWASP threats and compliance guardrails described later in this resource apply across all scenarios, but their relative priority and implementation specifics differ based on your scoping decision. Where relevant, each guardrail below is tagged with the scenario(s) where it is most critical.

The OWASP Top 10 for Agentic Applications for 2026 identified the following specific threats that must be mitigated for safe AI adoption:

- **ASI01: Agent Goal Hijack** - Attackers manipulate an agent's objectives, task selection, or decision pathways using techniques like prompt injection or deceptive tool outputs to redirect its autonomy toward harmful outcomes.
- **ASI02: Tool Misuse and Exploitation** - Agents misuse legitimate tools (e.g., deleting data, over-invoking costly APIs) due to prompt injection, misalignment, or unsafe delegation of authority.
- **ASI03: Identity and Privilege Abuse** - Exploiting the mismatch between user-centric identity and agentic design to escalate access, hijack privileges, or bypass controls through inherited or cached credentials.
- **ASI04: Agentic Supply Chain Vulnerabilities** - Malicious or tampered third-party components (models, tools, plugins, or MCP servers) introduce unsafe code or deceptive behaviors into the agent's execution chain.
- **ASI05: Unexpected Code Execution (RCE)** - Attackers exploit code-generation features or embedded tool access to execute unauthorized code (scripts, binaries, etc.) in real-time, bypassing traditional security controls.
- **ASI06: Memory & Context Poisoning** - Adversaries corrupt or seed an agent's stored context (conversation history, RAG stores, or embeddings) with misleading data to bias future reasoning and planning.
- **ASI07: Insecure Inter-Agent Communication** - Communication between agents (via APIs or message buses) that lacks proper authentication or integrity, allowing attackers to intercept, spoof, or manipulate intents.
- **ASI08: Cascading Failures** - A single fault (like a hallucination or poisoned memory) propagates and amplifies across multiple autonomous agents or systems, leading to widespread system-wide service failures.
- **ASI09: Human-Agent Trust Exploitation** - Exploiting human trust in agents (anthropomorphism) to influence user decisions, extract sensitive information, or bypass oversight.
- **ASI10: Rogue Agents** - Malicious or compromised agents that deviate from their intended function to act deceptively or parasitically within an ecosystem, resulting in a loss of behavioral integrity.

Risk Mitigation

To provide a robust mitigation plan for federal agencies, we must bridge the gap between high-level policy, like EO 14179, "Removing Barriers to American Leadership in Artificial Intelligence," and the technical reality of the OWASP Top 10 for Agentic Applications for 2026.

The Federal AI Guardrail Framework moves away from static compliance and toward continuous authority to operate (cATO), specifically for AI. (Note: The OWASP threats apply across all three AI implementation scenarios described above, but their severity and priority vary by scenario. Refer to the scoping matrix above to identify which threats are most critical for your deployment type.)

1. Risk: Lack of a Standardized Government-Wide Framework

The Federal Reality: Agencies are waiting for a "General AI FedRAMP", but cannot afford to stall mission delivery.

Mitigation Plan: The "Internal NIST AI RMF Overlay."

- Action: Immediately adopt the NIST AI Risk Management Framework (RMF 1.0) as the foundational baseline, but extend it with a custom "Agentic Security Overlay."
- Scoping: Most critical for Scenario 1 (Buying), where the agency has least visibility into the underlying model and must rely heavily on policy controls and AUPs to govern use.
- Technical Implementation: Map agency-specific NIST 800-53 controls to the OWASP Agentic Top 10. For example, use ASI01 (Goal Hijack) to define "System Characterization" boundaries.
- Deliverable: An agency-wide "AI Acceptable Use Policy" (AUP) that defines Trust Zones for LLMs (e.g., Public vs. Controlled Unclassified Information).

2. Risk: Algorithmic Bias, Transparency, and Explainability

The Federal Reality: Decisions affecting citizens (e.g., VA benefits or CDC health guidance) must be legally defensible.

Mitigation Plan: The "Inference Audit Trail" (IAT).

- Action: Implement Mandatory Explainability Requirements in the system prompt architecture.

- Scoping: Applies across all scenarios, but is most implementable in Scenarios 2 and 3 outlined above, where the agency controls the system prompt and model configuration.
- Technical Implementation: Utilize Chain-of-Thought (CoT) prompting with a hidden "Logic Log" that records the agent's internal reasoning steps before generating a final response. This mitigates ASI09 (Human-Agent Trust Exploitation) by providing an audit trail for why an AI reached a specific conclusion.
- Red Teaming: Conduct quarterly Bias Pressure Tests where agents are fed diverse demographic datasets to identify skew in benefits or service distribution.

3. Risk: Protection of Citizen Data Privacy

The Federal Reality: The leakage of personally identifiable information (PII), protected health information (PHI), and Controlled Unclassified Information (CUI) is a violation of the Privacy Act of 1974. Preventing CUI from being inadvertently disclosed to commercial LLMs is currently one of the highest priorities for federal security teams.

Mitigation Plan: "Zero-Trust Data Enclaves."

- Action: Move AI processing to the Data Plane rather than sending data to a centralized model.
- Scoping: Highest risk in Scenario 1 (Buying), where data may leave the agency perimeter. Most controllable in Scenario 3 (Owning), where data never leaves the agency's infrastructure.
- Technical Implementation: Use Privacy-Preserving RAG (Retrieval-Augmented Generation). Implement a local-to-the-data "PII Scrubber" service layer using cloud technologies that sits between the agency databases and the LLM.
- Data Minimization: Enforce ASI06 (Memory & Context Poisoning) mitigations by using Ephemeral Context Windows, ensuring the AI "forgets" specific citizen PII immediately after **the session concludes.**

4. Risk: Procurement and Lifecycle Management (Non-Compliant Deployments)

The Federal Reality: Traditional procurement buys "software," but AI is a "stochastic service" that changes daily.

Mitigation Plan: The "AIBOM" (AI Bill of Materials) Requirement.

- Action: Update all RFPs to require an AIBOM (as referenced in the OWASP document Appendix B).

- Scoping: Most critical for Scenario 1 (Buying) and Scenario 2 (Building), where third-party vendors supply core components. In Scenario 3 (Owning), the agency produces the AIBOM itself.
- Technical Implementation: Require vendors to disclose not just the model, but the training data lineage, safety tuning parameters, and third-party tools (MCP servers).
- Lifecycle Control: Implement a Kill-Switch Protocol (mitigating ASI10: Rogue Agents). If an AI's drift (monitored via cosine similarity or performance metrics) exceeds a threshold, the system automatically reverts to a "Safe State" or human-only mode.

5. Risk: Absence of a Unified AIaaS Platform

The Federal Reality: Every agency is reinventing the wheel, leading to "Shadow AI" usage.

Mitigation Plan: The "Sovereign AI Shared Service" (SAISS)

- Action: Build a multi-tenant Internal AI Gateway leveraging Cloud infrastructure that provides pre-authorized API access to models (Copilot/ChatGPT/Vertex AI/Bedrock) with government-specific guardrails.
- Scoping: Primarily addresses Scenario 1 (Buying) governance, but the Intent Gate architecture also serves as a control layer for Scenario 2 (Building) RAG pipelines.
- Technical Implementation: Build a "Policy Enforcement Middleware" (Intent Gate). This middleware intercepts every prompt to check for ASI02 (Tool Misuse) and ASI03 (Privilege Abuse) before the request ever reaches the LLM.
- Resource Sharing: Use Managed Workspaces where agencies share "Golden Prompt Templates" that have already been cleared by General Counsel for ethical compliance.

6. Risk: Scaling AI While Maintaining Public Trust

The Federal Reality: One "AI hallucination" in a public-facing service can set Federal AI adoption back by years.

Mitigation Plan: The "HITL (Human-in-the-Loop) Tiered Response"

- Action: Categorize all AI actions by Impact Level (Low, Medium, High).
- Scoping: Applies across all scenarios. In Scenario 2 and 3, agentic pipelines make the HITL tier especially important given the higher degree of autonomous action available to the system.

- Technical Implementation:
  - Low Impact (FAQ bot): Fully autonomous.
  - Medium Impact (Drafting a memo): Human-Reviewed.
  - High Impact (Decision making): Human-Authorized only.
- Trust Guard: Implement ASI08 (Cascading Failures) monitoring. If an AI agent attempts to trigger a secondary agent for a "High Impact" action, the system forces a hard-stop for manual human verification.

Key Considerations for AI Adoption

- **Scope** - Define the specific, measurable, achievable, relevant, and time-bound (SMART) goal for the AI application. Scope must also address the architectural question: is the agency buying a SaaS AI tool, building a RAG-based system on pre-trained models, or training/owning a custom model? Each scenario carries a fundamentally different risk profile (see AI Implementation Scoping section). Answering this question before designing controls prevents misallocated security resources and compliance gaps.
- **Components** - Identify and inventory all functional parts and user roles interacting with the AI system. This includes end users, data pipelines, the model itself, and any supporting infrastructure.
- **Resources** - Assess the required human capital across development, operation, security, and governance. This involves identifying skill gaps in areas such as machine learning operations (MLOps), AI ethics, and data science.
- **Risk Management** - Proactively identify and mitigate potential vulnerabilities. Forewarned is forearmed; consult the “Risk Mitigation” section below for recommendations on threat modeling, bias detection, and adversarial robustness.
- **Trade-offs** - Analyze the financial implications of varying levels of security and compliance rigor. Determine the optimal balance between investment in monitoring (e.g., real-time audit logs, enhanced data encryption) and the acceptable residual risk level, tailored to the application's scope and criticality of its components.

Mapping to Compliance Standards

Responsible AI adoption requires a structured compliance architecture. This section maps the Federal AI Guardrail Framework to three complementary standards:

- **ISO/IEC 42001:2023**, which governs the AI management lifecycle
- **ISO/IEC 27001:2022**, which secures the data and infrastructure supporting it
- **NIST AI RMF 1.0**, the federal benchmark previously required by Executive Order 14110 and now governed under EO 14179

While federal agencies primarily authorize systems using NIST SP 800-53 under FISMA, mapping to ISO/IEC 27001:2022 is critical for establishing supply chain security, evaluating commercial AI vendors, and ensuring interoperability with international mission partners. Together, these frameworks form a multi-layered approach that enables agencies to scale AI capabilities rapidly while maintaining public trust and adhering to legal mandates.

ISO/IEC 42001:2023

ISO/IEC 42001 focuses on the AI lifecycle. It is designed to manage the unique "stochastic" (unpredictable) nature of AI, which traditional software standards aren't equipped to handle. It is the world's first AI management system standard, providing valuable guidance for this rapidly changing field. It addresses the unique challenges posed by AI, such as ethical considerations, transparency, and continuous learning. For organizations, it sets out a structured way to manage risks and opportunities associated with AI, balancing innovation with governance.

Key Requirements for AI

The standard is structured into Clauses (management requirements) and Annex A (specific controls).

1. Context and Leadership (Clauses 4 & 5)

- **AI Policy:** Federal agencies must establish an AI policy that aligns with ethical principles and legal requirements.
- **Roles and Responsibilities:** Federal agencies must define who is responsible for AI oversight (e.g., an AI Officer or a cross-functional Governance Committee).

2. Planning and Risk Treatment (Clause 6)

- **AI Risk Assessment:** Unlike traditional IT risk, federal agencies must assess risks specific to AI, such as algorithmic bias, model drift, and lack of explainability.
- **AI Impact Assessment (AIIA):** Federal agencies must assess how the AI affects individuals and society (e.g., privacy, civil rights, or safety).

3. Operation and Lifecycle (Clause 8)

- **Lifecycle Management:** Requirements for the design, development, deployment, and retirement of AI systems.
- **Data for AI:** Ensuring data used for training and inference is high-quality, representative, and legally sourced.

4. Annex A: The Control Categories

Annex A contains 38 controls across 10 categories, including:

- **A.2 Internal Governance:** Establishing metrics for AI performance.
- **A.5 Resources for AI:** Managing the massive compute and specialized talent requirements.
- **A.8 Transparency & Explainability:** Ensuring the AI's "black box" can be audited.

Guardrails for ISO/IEC 42001:2023 Compliance

To get a federal agency into compliance, the agency needs to implement technical guardrails that act as the "enforcement layer" for the ISO 42001 policy. Below, we detail five technical guardrails, the requirements they map to, proposed solutions, and implementation guidance.

1. Implement a "Policy Enforcement Point" (PEP)

- **Requirement:** Clause 8 (Operational Control).
- **Scoping:** Most critical in Scenario 1 (Buying) and Scenario 2 (Building). In Scenario 3 (Owning), the PEP also monitors inference-time requests against the agency's own model.
- **Solution:** Build or procure an AI Gateway (Middleware). Every prompt and response passes through this gateway.
- **Implementation:** The gateway checks for PII, blocked keywords, or "Goal Hijacking" (referencing OWASP ASI01). If the prompt violates the AI policy, it is blocked before reaching the model.

2. Automated AI Impact Assessments (AIIA)

- **Requirement:** Clause 6.1 (AI Risk Assessment).
- **Scoping:** In Scenario 1, the AIBOM is required from the vendor. In Scenarios 2 and 3, the agency produces or contributes to the AIBOM as part of its own development pipeline.
- **Solution:** Use a Metadata Schema for every model deployment.
- **Implementation:** Before a developer can deploy a model to production (via automation), they must submit an AI Bill of Materials (AIBOM) that includes the data sources, intended use, and bias testing results. If these metadata fields aren't present, the continuous integration/continuous delivery (CI/CD) pipeline fails.

3. Continuous Monitoring for Model Drift (A.10 Monitoring)

- **Requirement:** Annex A.10 (System Monitoring).
- **Solution:** Implement Observability Sinks (GCP Cloud Monitoring / AWS CloudWatch).
- **Implementation:** Set up alerts for "Hallucination Variance." If the AI's responses begin to deviate significantly from a "Golden Dataset" (a set of pre-approved correct answers), the system triggers a manual human review, satisfying the ISO requirement for continuous oversight.

4. Explainability Logs (A.8 Transparency)

- **Requirement:** Annex A.8 (Transparency to stakeholders).
- **Solution:** Chain-of-Thought (CoT) persistence.
- **Implementation:** Force the model to generate its "internal reasoning" and store that reasoning in an immutable log (like BigQuery, Redshift, Athena). If a citizen or auditor asks, "Why was this decision made?", the Federal Agency can produce the log showing the AI's step-by-step logic.

5. Data Lineage and Provenance (A.6 Data for AI)

- **Requirement:** Annex A.6 (Data quality and origin).
- **Scoping:** Most operationally relevant in Scenario 3 (Owning) where the agency controls training data. In Scenario 2, DVC applies to the RAG corpus and fine-tuning datasets.
- **Solution:** Signed Data Sets.
- **Implementation:** Implement Data Version Control (DVC - an open-source tool for versioning datasets and ML models). Every version of the AI model must be cryptographically linked to the exact version of the training/fine-tuning dataset. This prevents "Memory Poisoning" (OWASP ASI06) and ensures auditability.

ISO/IEC 27001:2022

While ISO 42001 is the "Management System" for the AI itself, ISO/IEC 27001:2022 is the foundational standard for information security.

In the context of federal AI adoption, ISO 27001:2022 governs the data fed into the AI, the infrastructure that runs it, and the security of its outputs. The 2022 update is particularly important because it introduced new controls specifically designed for cloud services and digital footprints.

Key Requirements for AI

ISO 27001:2022 uses a "Process Approach" (Clauses 4-10) and a set of 93 controls (Annex A) categorized into four themes: organizational, people, physical, and technological.

1. Technological Controls (Annex A.8)
This is where the 2022 update hits AI adoption hardest.

- **A.8.10 Information Deletion:** Data must be deleted when no longer required. (Challenge: Managing data in LLM training caches or "memory".)
- **A.8.12 Data Leakage Prevention (DLP):** Preventing unauthorized disclosure of information. (Challenge: Preventing PII from being "leaked" through AI chat responses.)
- **A.8.28 Secure Coding:** Ensuring code is developed securely. (Challenge: AI-generated code often contains vulnerabilities.)

2. Context and Risk (Clauses 4 & 6)

- **Expansion of Scope:** The agency must include AI "information assets" (datasets, model weights, and API keys) in its asset register.
- **Risk Assessment:** Security leads must assess "emergent threats." You can't just protect servers; you must protect against **ASI01 (Goal Hijacking)** as a threat to information integrity.

3. Organizational Controls (Annex A.5)

- **A.5.7 Threat Intelligence:** Organizations must collect and analyze information about threats. (Requirement: Monitoring for new AI exploits.)
- **A.5.23 Information Security for use of Cloud Services:** Managing the risk of third-party AI providers (OpenAI, Anthropic, etc.).

Guardrails for ISO 27001:2022 Compliance

To move a federal agency toward ISO 27001 compliance for its AI projects, agencies can implement technical "checkpoints" that enforce these Annex A controls.

1. The "Input-Output" DLP (Data Loss Prevention) Filter

- **Requirement:** A.8.12 (Data Leakage Prevention).
- **Scoping:** Highest priority in Scenario 1 (Buying) where data crosses the perimeter to a commercial provider, and in Scenario 2 (Building) where citizen data is retrieved and passed to an external model.
- **Solution:** An automated PII/PHI scrubber.
- **Implementation:** Before a prompt reaches the LLM, it passes through a regex and localized (private) LLM-based scanner that masks sensitive data (e.g., "Social Security Number" becomes "[REDACTED]"). This ensures that sensitive data never leaves the Agency's "Security Perimeter," satisfying the Confidentiality requirement of 27001. This also prevents sensitive citizen data from being "learned" by the model itself.

2. Identity-Aware API Gateways

- **Requirement:** A.5.15 (Access Control) and A.5.18 (Access Rights).
- **Solution:** Identity-Aware Proxy (IAP) and Service Account scoping.
- **Implementation:** Instead of giving developers raw API keys (which are prone to ASI03: Identity Abuse), Agencies use short-lived OIDC (OpenID Connect) tokens. Agencies link every AI request to a specific Federal employee ID. If a "Rogue Agent" (ASI10) is detected, Agencies can immediately revoke that specific identity's access without affecting the rest of the system or its users.

3. AI-Generated Code Sandboxing

- **Requirement:** A.8.28 (Secure Coding) and A.8.25 (Secure Development Lifecycle).
- **Solution:** Automated SAST (Static Application Security Testing) for AI.
- **Implementation:** If an Agency uses an AI "Coding Assistant" (like GitHub Copilot or Vertex AI), any code generated must be automatically scanned by a tool like Snyk or SonarQube before it is allowed into a Pull Request. This prevents the introduction of ASI05 (Unexpected Code Execution) vulnerabilities.

4. Immutable Logging and SIEM Integration

- **Requirement:** A.8.16 (Monitoring Activities).
- **Solution:** Log Sinks to BigQuery/S3/Splunk with mTLS.
- **Implementation:** Federal Agencies must maintain logs that prove who accessed what data. Every AI interaction (Input + Output + Metadata) should be streamed to a secure, immutable log. Agencies integrate these logs into the Agency's Security Operations Center (SOC) so that "Anomalous AI Behavior" triggers the same alerts as a standard network intrusion.

5. Threat Intelligence Feeds

- **Requirement:** A.5.7 (Threat Intelligence).
- **Solution:** Automated Vulnerability Feed.
- **Implementation:** Agencies configure their security tools to ingest the OWASP Agentic AI Top 10 and the MITRE Adversarial Threat Landscape for AI Systems (ATLAS) database to proactively defend against emerging exploits.

NIST AI RMF 1.0

Originally mandated by Executive Order 14110 (revoked January 20, 2025), the NIST AI Risk Management Framework remains a foundational federal benchmark under EO 14179. It focuses on the socio-technical risks of AI and divides activities into four functions: map, measure, manage, and govern.

Map (Contextual Risk)

The NIST AI RMF defines Map as the process of establishing context - identifying the AI system's purpose, potential harms, and the organizational and societal environment in which it will operate. Agencies seeking to operationalize this function within a federal environment may consider extending the Map baseline with agency-specific risk boundaries, prompt routing controls, and formal acceptable use policies that define how AI systems interact with different data classifications and user populations.

Recommended Implementation Actions:

- Adopt the NIST AI Risk Management Framework (RMF) 1.0 as the foundational baseline, extending it with an "Agentic Security Overlay."
- Map agency-specific NIST 800-53 controls to the OWASP Agentic Top 10, for example, using ASI01 (Goal Hijack) to define "System Characterization" boundaries.
- The Map function should explicitly incorporate the AI Implementation Scoping determination (Buying / Building / Owning) as a foundational context-setting step. The Trust Zone definitions in the AUP should reflect the risk profile of the agency's chosen implementation scenario.
- Deliver an agency-wide "AI Acceptable Use Policy" (AUP) that formally defines Trust Zones for LLMs (e.g., Public vs. Controlled Unclassified Information).
- Implement a Policy Enforcement Point (PEP), an AI Gateway (Middleware) that intercepts every prompt to check for PII, blocked keywords, or "Goal Hijacking" (ASI01), blocking the request if policy is violated.
- Use a Policy Enforcement Middleware (Intent Gate) to check for ASI02 (Tool Misuse) and ASI03 (Privilege Abuse) before a request reaches the LLM.

Measure (Test, Evaluation, Verification, and Validation (TEVV))

The NIST AI RMF defines Measure as the application of TEVV processes to assess, track, and document AI risks and trustworthiness characteristics throughout the system lifecycle. Agencies looking to build a continuous and auditable TEVV capability may consider implementing automated assessment pipelines, standardized metadata requirements at deployment, and structured bias testing protocols to ensure risk measurement keeps pace with system changes.

Recommended Implementation Actions:

- Conduct quarterly Bias Pressure Tests where agents are fed diverse demographic datasets to identify skew in benefits or service distribution.
- Implement Automated AI Impact Assessments (AIIA) by requiring a Metadata Schema for every model deployment.
- As part of the deployment pipeline, a developer must submit an AI Bill of Materials (AIBOM) detailing data sources, intended use, and bias testing results; the CI/CD pipeline fails if these metadata fields are absent.
- Use Automated SAST (Static Application Security Testing) for any AI-generated code to prevent the introduction of ASI05 (Unexpected Code Execution) vulnerabilities.

Manage (Active Risk Control)

The NIST AI RMF defines Manage as the active prioritization and treatment of identified risks, including allocating resources and responses proportional to an AI system's assessed risk level and potential impact. Agencies seeking to move beyond reactive risk response may consider implementing automated monitoring, defined intervention thresholds, and tiered human-in-the-loop controls that allow systems to be contained or reverted when drift or anomalous behavior is detected.

Recommended Implementation Actions:

- Implement Continuous Monitoring for Model Drift using Observability Sinks (e.g., GCP Cloud Monitoring / AWS CloudWatch).
- Set up alerts for "Hallucination Variance": if the AI's responses deviate significantly from a "Golden Dataset" of pre-approved correct answers, the system triggers a manual human review.
- Implement a Kill-Switch Protocol (mitigating ASI10: Rogue Agents) that automatically reverts the system to a "Safe State" or human-only mode if the AI's drift exceeds a threshold.
- Implement the "Zero-Trust Data Enclaves" mitigation plan, which moves AI processing to the Data Plane rather than sending data to a centralized model.
- Utilize Ephemeral Context Windows to enforce ASI06 (Memory & Context Poisoning) mitigations, ensuring the AI "forgets" specific citizen PII immediately after the session concludes.
- Enforce a Human-in-the-Loop (HITL) Tiered Response: for High Impact actions (like decision-making), the action must be Human-Authorized only.

Govern (Transparency)

The NIST AI RMF defines Govern as the cross-cutting function that establishes organizational accountability, culture, and processes to support responsible AI risk management across all other functions.

Agencies working to meet transparency and auditability requirements may consider mandating documentation of data lineage, model provenance, and internal reasoning at the point of deployment - ensuring that accounta

Recommended Implementation Actions:

- Implement the "Inference Audit Trail" (IAT) by utilizing Chain-of-Thought (CoT) prompting with a hidden "Logic Log" that records the agent's internal reasoning steps before generating a final response, which mitigates ASI09 (Human-Agent Trust Exploitation).
- Store the AI's "internal reasoning" in an immutable log (CoT persistence) (e.g., BigQuery, Redshift, Athena) to provide an audit trail for decisions.
- Require vendors to disclose the training data lineage, safety tuning parameters, and third-party tools (MCP (Model Context Protocol) servers) as part of the AIBOM.
- Implement Data Version Control (DVC) to cryptographically link every version of the AI model to the exact version of the training/fine-tuning dataset, which ensures auditability and prevents ASI06 (Memory Poisoning).bility is built into the system from the start rather than applied after the fact.