Practitioner White Paper

AI Security Integration Framework (AISIF)

A Proposed Framework to Connect Seven AI Security Frameworks into One Working Program

by

Olasupo (Ola) Lawal, *MSc, MBA, CISSP, ISO 27001 Senior Lead Implementer*
www.linkedin.com/in/ola-lawal
E-mail: ola@olalawal.com

Version 1.0 | 2026 | Released under Creative Commons CC BY 4.0
Designed for: CISOs, Security Architects, AI Security Engineers, GRC Leaders, Audit and Compliance Teams

What this document is

A practitioner guide to the AI Security Integration Framework (AISIF), a unified recursive model for governing engineering and operating secure AI systems. This framework is a five-layer architecture that maps how seven principal AI security frameworks can be leveraged to work together across five layers: Governance, Data, Model, Application, and Operations. It includes a maturity model that practitioners can use to determine the current maturity of an organization's AI security program, while identifying the highest priority improvements. It also proposes a structured control taxonomy with mappings to internationally recognized standards.

What you will be able to do after reading it

- Understand which of the seven frameworks applies to which part of your AI security program — and stop treating them as separate work streams
- Identify the layer or layers where your program is most exposed using the AISIF Maturity Model
- Map your existing controls to ISO 27001, NIST AI RMF, and ISO 42001 using the standards crosswalk
- Use the three case studies to run a gap analysis against real attack scenarios
- Walk into your next architecture review, threat model, or audit with a unified framework and a structured artefact

1. The Problem: Seven Frameworks, No Integration

> **Most organizations deploying AI systems are not short on frameworks but are rather short on integration.**

If you are building or maturing a GenAI security program, you have likely encountered at least a few of the seven frameworks covered in this document. You would have also observed that each one answers a different question, while none of them answers all questions or threats. That is not a design flaw; it reflects the genuine complexity of securing AI systems, which span adversarial threat surfaces, application vulnerabilities, data pipelines, cloud architecture, and regulatory obligations simultaneously. Your AI security framework bookmarks probably include NIST AI RMF, OWASP LLM Top 10, and MITRE ATLAS. You may have run an AWS Scoping Matrix exercise, read the Google SAIF documentation, or had a regulatory conversation about the EU AI Act. You may even have an ISO 27001 ISMS that predates your AI work by several years.

When a prompt injection attack succeeds in a RAG-based application, a training data poisoning scenario plays out across a monthly retraining cycle, or a model inversion attack extracts sensitive data from a fine-tuned LLM, the common finding is not that the organization lacked a framework. It is that the organization applied its frameworks episodically, in isolation, and without a model for how they connect.

The OWASP LLM Top 10 might have been referenced during the architecture review. NIST AI RMF may have been cited in the governance document. MITRE ATLAS might have been the basis for the red team exercise that ran once and was never repeated. The operating model that makes all seven works together was never considered or built and nobody mapped the resulting controls to the standards against which the program was audited. This is a common problem that AISF solves.

The Seven Frameworks: What Each One Does and Where It Stops

To understand the integration gap, we need to start with an understanding of each framework, what they are designed to do, and, equally important, their respective limitations.

MITRE ATLAS (Adversarial Threat Intelligence for AI)

MITRE ATLAS is the framework you reach for when you need to understand how an adversary goes after an AI system. It extends the MITRE ATT&CK model into the machine learning domain, cataloguing real-world adversarial tactics across the full attack lifecycle: reconnaissance, resource development, initial access, model evasion, data poisoning, and model theft. If your threat modeling

practice stops at the application perimeter and does not account for ML specific attack vectors, ATLAS is your gap.

For security teams already comfortable with ATT&CK-based threat modeling, the learning curve is low. The framework is directly actionable for red team exercises and adversarial simulation. Its limitation is its scope, as it is an attack taxonomy, not a governance or control framework. While it will tell you what can go wrong, pairing it with Google SAIF or NIST AI RMF will tell you what to do about it.

OWASP Top 10 for LLM Applications (Application-Layer Risk Reference)

OWASP's LLM Top 10 is the most widely cited reference for application layer Gen AI risk, and for very good reasons. It is practitioner-facing and grounded in observed failures rather than theoretical attack models. The ten risks include prompt injection, insecure output handling, training data poisoning, model denial of service, excessive agency, and insecure plugin design, among others.

If your AppSec team is reviewing LLM-integrated applications and needs a structured risk checklist, this is the starting point. Its limitation is that it does not extend into enterprise governance, data pipeline security, or regulatory compliance. Think of it as the application security layer of a broader program, which is necessary, but is not sufficient on its own.

AWS Generative AI Security Scoping Matrix (Clarifying Accountability Before Controls)

To design effective controls for an AI system, you need to know what you own and operate. That is the problem the AWS Scoping Matrix solves. It maps security responsibilities across five deployment archetypes, from consuming a fully managed third-party model to training and operating your own foundation model, while identifying where accountability lies at each layer of the stack.

This framework is particularly valuable for organizations that have grown their AI portfolio quickly and have not formally documented and assigned cyber security roles and responsibilities. Although it does not prescribe controls, it is a prerequisite for doing so coherently. Running through this exercise before your next architecture review will surface accountability gaps that most teams have not explicitly addressed.

Google Secure AI Framework (SAIF) (Secure-by-Design for AI Systems)

SAIF reflects Google's position as both a leading AI developer and a security engineering organization. It advocates for embedding security controls across the full AI development lifecycle rather than applying them after the fact. The framework covers six core elements: expanding the security perimeter to include the ML supply chain, extending detection and response to AI-specific

signals, automating defenses at scale, harmonizing platform-level controls, adapting controls as threats evolve, and contextualizing AI risk within broader organizational risk.

For security engineers and platform teams, SAIF provides the most technically grounded guidance of the seven frameworks. Its weakness is that it is not oriented toward compliance. If your primary driver is regulatory alignment, you will need to layer NIST AI RMF and ISO 42001 on top of it.

Databricks AI Security Framework (Security Starts in the Data Plane)

Most security frameworks treat the model as the primary asset to protect. The Databricks AI Security Framework starts from a different premise, that a model is only as trustworthy as the data that trained it and the pipelines that feed it. If your data estate is poorly governed, your AI system is compromised before any adversary acts.

The framework addresses security across the full lake house architecture, from data ingestion and transformation through feature engineering, model training, deployment, and inference monitoring. Key control areas include:

- Data governance, classification, and lineage tracking
- Access control and isolation across the data and ML platform
- Model lifecycle security from training artifacts to serving endpoints
- Runtime monitoring for data drift, misuse patterns, and anomalous outputs
- Unified security posture across data, analytics, and AI workloads

For organizations running AI workloads on Databricks or similar lake house platforms, this framework addresses a set of risks that the other six largely ignore. Even if you are not on Databricks, the conceptual emphasis on data-plane security is applicable to any organization with significant data pipeline complexity.

EU AI Act (Regulatory Obligations and Risk Classification)

The EU AI Act is not a security framework in a technical sense. It is binding law that establishes risk-based obligations for AI systems deployed in or affecting the European Union. It classifies AI systems into four risk tiers (unacceptable, high, limited, and minimal risk) and imposes corresponding requirements for transparency, human oversight, technical robustness, data governance, and conformity assessment.

For practitioners, the most operationally significant tier is high-risk, which includes AI systems used in critical infrastructure, employment decisions, credit scoring, law enforcement, and several other domains. High-risk systems must comply with detailed technical and governance requirements, and

organizations must maintain documentation sufficient to demonstrate conformity. ISO/IEC 42001 is increasingly referenced as a conformity mechanism. If your organization operates in scope, compliance is not discretionary, and the technical controls required map directly to several of the other frameworks in this document.

NIST AI Risk Management Framework — Enterprise Governance for AI Risk

The NIST AI RMF is the most comprehensive governance framework in this set, and for many organizations it is the right anchor for an enterprise AI security program. Its four core functions Govern, Map, Measure, Manage provides a structured, iterative approach to identifying and evaluating AI risks, while maintaining ongoing accountability for how they are addressed.

There are a few things worth understanding about the NIST AI RMF. First, it is deliberately non-prescriptive, as it does not tell you exactly which controls to implement. Second, it tells you how to build and operate a program that can make those decisions systematically and demonstrate accountability for them. This is by design as the framework is intended to be applicable across industries, AI system types, and risk appetites. The practical implication is that you will need to pair it with more prescriptive frameworks (MITRE ATLAS for threat modeling, OWASP LLM top 10 for application risks, Google SAIF or Databricks for technical controls, to operationalize it fully.

Where they Converge

Despite their different origins and audiences, all seven frameworks share a set of foundational assumptions that are worth mentioning, because they represent the areas where the evidence is strongest and the practitioner consensus is clearest:

- Data integrity is non-negotiable: Every framework that touches data treats its integrity as a prerequisite for trustworthy AI, not an optional enhancement.
- Monitoring cannot be a deployment-phase afterthought: Continuous evaluation of model behavior, data drift, and system outputs is required across the board, not just at initial release.
- AI systems are socio-technical: Security controls that only address the technical stack (models, infrastructure, APIs), without addressing the organizational and human factors will have structural blind spots.
- Lifecycle coverage is required: Point-in-time security assessments are insufficient. Controls need to span design, development, deployment, and decommission.

These points of convergence are useful because they outline where to concentrate effort regardless of which frameworks an organization has formally adopted. If you are not confident in your data integrity controls, monitoring practices, and lifecycle governance, those are the right places to start, before optimizing for any framework's specific requirements.

Where they Diverge

The more useful insight, however, is where the frameworks separate, because that tells us which ones to is best suited for which problems. The clearest dividing line is between frameworks oriented toward attack and failure, and those oriented toward construction, governance, and accountability:

- **Attack and failure analysis:** MITRE ATLAS, OWASP LLM Top 10
- **Secure construction and operations:** Google SAIF, Databricks AI Security Framework
- **Accountability and ownership:** AWS Generative AI Security Scoping Matrix
- **Governance and regulatory justification:** NIST AI RMF, EU AI Act

If an organization is pulling exclusively from NIST AI RMF and OWASP LLM Top 10, it will likely have reasonable governance documentation and application layer risk awareness, but insufficient adversarial threat modeling and weak data-plane controls. If they are heavy on MITRE ATLAS and Google SAIF, they probably have strong technical controls but may struggle to demonstrate regulatory alignment or executive-level accountability. The goal of this document is to help organizations and practitioners identify and close those asymmetries.

The Databricks framework occupies a unique position worth calling out. It is the only framework that treats the data pipeline, not the model, not the application, not the governance process as the primary attack surface. For organizations with significant data engineering complexity, this perspective addresses a real and frequently underinvested risk area.

Table 1: Strengths and Limitations of Common AI Security Frameworks

Framework	Best Used For	Main Audience	AISIF Layer(s)	Strength	Limitations
MITRE ATLAS	Red team planning, ML-specific threat modeling	Security, red teams	Model	Real adversarial TTPs for ML systems	Does not cover governance, compliance, and controls
OWASP LLM Top 10	Application security reviews, code reviews	Developers, AppSec	Application	Concrete, developer-ready risk catalogue	Does not cover data pipelines, enterprise governance
AWS Scoping Matrix	Cloud AI accountability mapping, vendor onboarding	Cloud architects	Governance	Clarifies who owns what before controls are built	Does not prescribe actual controls
Google SAIF	Secure-by-design engineering, platform controls	Security engineers	Model, Application, Operations	Best technical layered defense guidance	Lacks guidance on regulatory compliance demonstration
Databricks AI Security	Data pipeline and lake house security	Data/ML engineers	Data, Operations	Operationalizes data-plane controls no others addresses	Ignores the application layer, and governance
EU AI Act	Risk tier classification, regulatory compliance	Legal, risk, product	Governance	Provides binding legal authority, with clear obligations	Lacks technical implementation guidance
NIST AI RMF	Enterprise AI risk governance Program design	Executives, risk leaders	Governance	Holistic lifecycle governance model	Specific control selection without supplements

Table 1 above shows two important things about these AI security frameworks.

First, the seven frameworks are genuinely complementary, as there is no meaningful redundancy between them. Second, each one has a real limitation that at least one other framework compensates for. MITRE ATLAS tells you how attacks happen but does not provide guidance on what to do about them. OWASP LLM Top 10 tells you what application-layer failures look like but not how to govern the program that prevents them. NIST AI RMF tells you how to organize a governance program but does not articulate technical controls to implement. The only way to get full coverage is to connect these frameworks together.

The Five Failure Patterns

Security programs that engage with these frameworks individually tend to fail in the same five ways. Recognizing which pattern applies to your program is the first diagnostic step AISIF supports.

Data pipeline security is often underinvested relative to model security

This seems to be the most common gap in many organizations. Many organizations invest in model evaluation, API security, and application-layer controls while leaving the data ingestion, transformation, and feature engineering pipelines that feed those models largely uncontrolled. If your training data can be tampered with, without detection, until a downstream model failure brings it to light, this is your gap.

Accountability mapping is assumed rather than documented

Organizations deploying AI through third-party providers do not frequently document which security controls are the provider's responsibility and which are theirs. The AWS Scoping Matrix provides guidance on this. Without it, roles and responsibilities in a cloud environment defaults to assumption, which will likely fail during incidents.

Adversarial testing is generic rather than ML-specific

Traditional penetration and application security testing does not cover ML specific attack vectors, unless specifically included. Red team exercises that do not reference MITRE ATLAS tactics will likely miss model evasion, training data poisoning, and model extraction entirely. If your last red team engagement had no ML specific scope, (or AI application testing does not generally include it), this is your gap.

Monitoring is often front-loaded at deployment and declines thereafter

In many organizations, pre-deployment testing typically receives significant investment. However, post-deployment monitoring of model behavior, output quality, and data drift typically receives far less attention or investment. The risk profile of a deployed AI system only begins at deployment, and it changes continuously. Front-loaded testing produces a point-in-time assurance that becomes ineffective over time, with little to no follow up.

Regulatory compliance and security posture are treated as the same thing

Meeting EU AI Act requirements for a high-risk system demonstrates operational maturity in documented governance and conformity assessment processes. As compliance does not necessarily equal security, checking off the compliance boxes does not mean that the system is secure against MITRE ATLAS attack tactics or OWASP LLM Top 10 application vulnerabilities. Compliance and security are related but distinct objectives. Organizations that equate them will likely have documented risks they have not mitigated.

2. The AISIF Architecture

AISIF organizes AI security controls across five layers. Each layer addresses a distinct portion of the AI system's security surface, maps to a specific set of frameworks, and has a defined relationship to the layers above and below it. The layers are not independent silos but are connected through a recursive feedback loop. This feedback loop is the major distinguishing architectural feature of AISIF.

Table 2: AISIF Five-Layer Architecture with Recursive Feedback Loop

Layer		Layer Name	What It Secures	Framework Alignment	If You Skip It...
CONTINUOUS MONITORING & FEEDBACK LOOP	1	**Governance**	Risk management cycle, AI policy, accountability, regulatory compliance, ethics review	NIST AI RMF Govern ISO 42001 Cl. 4–10 EU AI Act	Controls have no owner and no feedback path. Governance becomes a document, not a program.
	2	**Data**	Data lineage, provenance, access control, pipeline integrity, masking, retention	Databricks AI Security ISO 27001 A.8–A.9 ISO 42001 Annex A.6	Your model is only as trustworthy as your data. An ungoverned data estate will likely lead to a compromised model.
	3	**Model**	Training validation, adversarial robustness, model provenance, supply chain, API access	MITRE ATLAS Google SAIF NIST AI RMF Map/Measure	ML specific attack vectors including poisoning, evasion, model theft will be invisible to your security team.
	4	**Application**	Prompt injection defense, output filtering, plugin permissions, RAG sanitization, transparency	OWASP LLM Top 10 Google SAIF AWS Scoping Matrix	The most immediately accessible and exploitable attack surface in your AI system is completely undefended.
	5	**Operations**	Continuous monitoring, drift detection, incident response, adversarial log review, feedback loop	Databricks NIST AI RMF Manage ISO 27001 A.8.15–16	You will likely find out about failures from users, not your monitoring. Risk profile drifts invisibly post-deployment.

> The recursive feedback loop is what makes AISIF different from a framework catalogue. Operations findings feedback to Governance. Governance updates controls across all layers. The cycle runs continuously, not once at deployment.

Understanding the Recursive Feedback Loop

Every layer in AISIF produces output, that becomes the input for the next layer down. Governance defines the policy and risk tolerance that data controls enforce. Data integrity determines what the model trains on. Model robustness determines what the application can safely expose. Application outputs are what operations monitors.

But the flow is not one-directional, as operations layer is not a terminal layer. It is the origin of a feedback signal. Operational monitoring produces anomalies, incidents, drift measurements, and adversarial detection events. Important outputs that are evaluated in the governance layer's risk management cycle. If these outputs indicate that current controls are insufficient, governance triggers updates at the appropriate lower layer. Those updates will then be implemented, and the cycle continues.

This recursive property reflects the empirical reality of deployed AI systems, their threat landscape, capability boundaries, and regulatory context which does not stay fixed. A model that was secure at deployment may become a liability six months later through data drift, capability updates, or the emergence of new adversarial techniques. A governance program that does not have a feedback mechanism from operations will eventually become stale, accurately describing the system's risk posture at a point in time that has passed.

In practical terms, running the AISIF feedback loop means an organization is monitoring findings, reaches it’s security governance cycle on a defined or periodic schedule, and that governance cycle has the authority and processes to update controls in response. If your organization cannot describe that cycle, i.e., who receives the monitoring findings, at what frequency, and what authority triggers a control update, then the feedback loop does not exist in your program.

ASIF Layers: What You Need in Place

Layer 1 — Governance

The Governance layer establishes the organizational structure within which all other security decisions are made. Without it, lower-layer controls have no owner, no escalation path, and no adaptation mechanisms. Three key questions tell you whether your Governance layer is functional:

- Do you have a documented AI use policy that covers this system and has been reviewed in the last 12 months?
- Is there a named individual with decision authority over this system's security and risk posture?
- Have you completed an AI risk assessment that is specific to this system — not a generic IT risk assessment?

If the responses to any of these three questions is 'no', the Governance layer is not yet functional, and any controls built on it are built on an insecure foundation. Address these before investing further in technical controls.

Layer 2 — Data

The Data layer is the most commonly underinvested layer in AI security programs and the one with the longest-range consequences. A training data poisoning attack that is not detected in the Data layer can take several retraining cycles to manifest as a visible model failure, months after the initial compromise. The minimum viable Data layer controls are:

- End-to-end data lineage: can you trace every record in your training set back to its source?
- Are Access controls on training datasets and feature stores, enforced at the system level, not just in policy?
- Is statistical distribution monitoring conducted on data batches before they are used for training or inference?
- Is sensitive data masked or pseudonymized before it enters the training pipeline?

Layer 3 — Model

The Model layer is where MITRE ATLAS-informed security work belongs. Most organisations that run red team exercises do so at the application layer, testing prompt injections, output filtering, and API access. The model layer requires different testing. Adversarial example crafting, model evasion evaluation, model extraction attempts, and supply chain validation of any pre-trained model artifacts, ae important source testing in this layer to mention a few. Key questions are:

- When was the last adversarial test conducted that specifically covered ML specific attack vectors from the MITRE ATLAS taxonomy?
- Is the provenance of every model artifact in your environment documented, including pre-trained base models?
- Are inference API endpoints securely authenticated with the same rigor as other sensitive system APIs?

Layer 4 — Application

The Application layer is the most immediately exploitable surface for most deployed AI systems, and the OWASP LLM Top 10 is the right starting reference. The highest-priority controls are prompt injection defence (LLM01), output filtering and PII detection (LLM02), as well as plugin/tool permission scoping (LLM07/08). If you have a RAG system, indirect prompt injection, where adversarial instructions are embedded in retrieved documents rather than user input, requires explicit source validation controls that standard input sanitisation will not catch.

Layer 5 — Operations

The Operations layer is where security programs most commonly fail after an initially strong deployment. The pattern is consistent, as significant pre-deployment testing investment, followed by declining monitoring effort as the system becomes operational.

The Operations layer requires:

- Continuous inference logging with anomaly alerting, not periodic review.
- Data drift monitoring with defined thresholds that trigger a governance review.
- An AI-specific incident response plan with runbooks calibrated for ML failure modes (model rollback, training data quarantine, prompt injection forensics).
- A defined schedule for adversarial log review that looks for systematic boundary-testing queries and extraction patterns.

3. Case Studies: AISIF Applied to Real Attack Scenarios

The following five scenarios illustrate how AISIF's layer model can be used to analyze AI security failures. Case Studies A, B, and C are illustrative composites; Case Studies D and E analyze publicly disclosed incidents, providing the empirical grounding identified as future work in earlier versions of this paper. Each scenario involves a failure that crosses layer boundaries, where a gap in one layer enables a failure in another. Although a single-framework analysis would have identified part of the problem, AISIF makes the cross-layer failure visible and actionable — and the empirical cases demonstrate that the seam runs in both directions: A–C show AI-specific failures that traditional frameworks miss, while D and E show conventional failures that AI-specific frameworks miss.

Table 3: Prompt Injection Case Study

Case Study A — **Prompt Injection in a Healthcare RAG Chatbot** | Healthcare	
Attack vector	An adversary embeds prompt injection payloads in external documents uploaded to an EHR system. The RAG pipeline retrieves these documents as context and passes them to the LLM without sanitisation, allowing the injected instructions to override the system prompt and exfiltrate patient data.
AISIF layers implicated	• Application Layer (primary failure) • Data Layer (secondary failure)
Failure mode	The RAG pipeline treated retrieved content as trusted context. There was no hierarchy enforcement instruction to separate system prompts from retrieved data, and no output monitoring detected the data exfiltration pattern.
AISIF mitigation	• Application Layer: OWASP LLM01 indirect injection controls to ensure there is structural separation of retrieved context from system instructions. • Data Layer: Content validation at EHR ingestion stage to prevent adversarial payloads entering the retrieval corpus.
Standards activated	• ISO 27001 A.8.28, A.8.29 • NIST AI RMF MANAGE 2.2 • ISO 42001 Cl. 8.5

Table 4: Training Data Poisoning Case Study

Case Study B — Training Data Poisoning in an Enterprise ML Pipeline | Financial Services	
Attack vector	An adversary with access to a legacy transaction database introduces carefully crafted records over three months. The perturbations are designed to shift the model's decision boundary in ways that underestimate the risk of a specific fraud class. The effect is invisible in standard model evaluation metrics but manifests as elevated fraud losses over the six months following retraining.

AISIF layers implicated	• Data Layer (primary failure) • Model Layer (secondary failure)
Failure mode	There was no distribution monitoring on upstream data batches, no hash validation of training data inputs, and no pre-deployment adversarial validation of the model's decision boundary. Three monthly retraining cycles passed before the poisoning was detectable.
AISIF mitigation	• Data Layer: Implement data bricks style pipeline integrity controls, including hash-based batch validation and statistical distribution monitoring that would have flagged the anomalous records. • Model Layer: ATLAS-informed decision boundary analysis prior to each deployment.
Standards activated	• ISO 27001 A.8.9 • NIST AI RMF MAP 3.5 / MEASURE 2.5 • ISO 42001 Cl. 8.3.2

Table 5: Model Inversion Attack Case Study

Case Study C — **Model Inversion Attack on a Fine-Tuned Financial LLM** | Financial Services	
Attack vector	A fine-tuned LLM trained on proprietary customer financial data is exposed via a mobile application API. An attacker submits systematic query sequences designed to mimic normal user behaviour. Over several weeks, partial reconstruction of customer financial profiles from the model's outputs becomes possible.
AISIF layers implicated	• Model Layer (primary failure) • Operations Layer (secondary failure)
Failure mode	The model was fine-tuned without differential privacy protections. API rate limiting was insufficient to detect systematic extraction patterns. No behavioural baseline existed for authenticated user query patterns, so the anomalous sequences were not detected and flagged.
AISIF mitigation	• Model Layer: Differential privacy should be applied at fine-tuning stage. • Operations Layer: Inference query pattern monitoring with behavioral baselines, deviation consistent with systematic extraction should trigger review.
Standards activated	• ISO 27001 A.5.15, A.8.16 • NIST AI RMF MEASURE 2.6 / MANAGE 4.1 • ISO 42001 Cl. 9.1

> **Pattern across all three:** The failure always crosses layer boundaries. In Case A, a Data Layer gap enabled an Application Layer exploit. In Case B, a Data Layer absence let a Model Layer failure run undetected. In Case C, a Model Layer design decision created an Operations Layer blind spot. If your security program operates in single-layer mode, you detect part of each failure, but certainly not all of it.

Table 5A: DeepSeek Infrastructure Exposure Case Study (Empirical)

Case Study D — **Infrastructure Exposure: DeepSeek ClickHouse Database** | AI Model Provider (empirical, January 2025)	
Attack vector	A publicly accessible ClickHouse database on internet-facing DeepSeek subdomains (ports 8123 and 9000), reachable without any authentication. The HTTP interface permitted arbitrary SQL execution from a browser, exposing more than one million log-stream entries containing plaintext chat history, API secrets, and backend details — and permitting full control of database operations. Wiz Research identified the exposure within minutes of an external posture assessment, days after DeepSeek's models went viral; no adversarial AI technique was required or used.
AISIF layers implicated	Operations Layer (primary failure) — unauthenticated, unsegmented, internet-exposed datastore; no attack-surface monitoring; detection by external disclosure rather than internal telemetry. Data Layer (primary amplifier) — inference logs treated as operational exhaust rather than classified user data; chat content and secrets persisted in plaintext with no minimisation at ingestion. Governance Layer (root cause) — no pre-production exposure gate operating at deployment velocity. Model Layer — not attacked; latent upward risk propagation only (exposed secrets enable credentialed abuse of inference endpoints).
Failure mode	Governance-velocity failure manifesting as Operations-layer control absence, with Data-layer classification failure as the amplifier. No AI-specific attack surface was involved. Single-framework analysis misses this entirely: no OWASP LLM Top 10 category fits an unauthenticated ancillary datastore, and no ATLAS technique was employed — the preventing controls are ISO/IEC 27001 Annex A fundamentals. In an AI system, the telemetry is the user data: log pipelines carry the same sensitivity class as the primary application database.
AISIF mitigation	Operations Layer: authentication and network isolation on all AI-adjacent datastores, including analytics and observability stores; continuous external attack-surface monitoring so exposure is detected internally before it is detected externally. Data Layer: classify inference logs at the sensitivity of their contents; redact chat content and secrets at log ingestion; minimisation by default for derived stores. Governance Layer: a mandatory pre-production exposure-review gate whose cadence is bound to deployment cadence, not a periodic review calendar.
Standards activated	ISO/IEC 27001:2022 Annex A (access control, secure configuration, network security, asset management) NIST AI RMF Govern and Manage functions; NIST AI 600-1 data-privacy and information-security actions CSA AI Controls Matrix infrastructure and data-security domains; OWASP LLM02 (downstream, partial)

Table 5B: Snowflake Identity Compromise Case Study (Empirical)

Case Study E — **Identity Compromise of the AI Data Estate: The Snowflake Customer Campaign** | Cloud Data Platform Customers (empirical, 2024)	
Attack vector	Threat actor UNC5537 accessed approximately 165 Snowflake customer tenants using valid customer credentials harvested by commodity infostealer malware, with the earliest harvested credential dating to November 2020. Access succeeded because the affected accounts lacked multi-factor authentication, credentials had never been rotated, and tenants had no network allow-lists. Many infections resided on third-party contractor endpoints holding elevated privileges across multiple clients. Data was exfiltrated at scale and victims extorted; confirmed victims included Ticketmaster/Live Nation, Santander, AT&T, Advance Auto Parts, and Neiman Marcus. Snowflake's own platform was not breached — and no novel technique was required.
AISIF layers implicated	Governance Layer (root cause) — shared-responsibility ambiguity: platform-available controls (MFA, network policy) were not mandated by customer standards; contractor access was not held to the customers' own identity and endpoint requirements. Operations Layer (primary failure locus, customer side) — single-factor authentication, multi-year credential staleness, no network policy, no behavioural detection on authentication or bulk egress. Data Layer (asset at risk) — the compromised tenants constitute the analytical data estate from which AI training corpora, feature stores, and retrieval indexes are drawn. Application and Model Layers — latent: pipeline service identities share the same weaknesses; write access to source data is a poisoning capability against every downstream model.
Failure mode	Shared-responsibility governance failure manifesting as identity-trust absence at the Operations layer, with third-party endpoint compromise as the initial access vector. The relevant adversary technique catalogue is MITRE ATT&CK (Valid Accounts, T1078) — not ATLAS — and AI-specific frameworks are silent on single-factor authentication for a data warehouse. The operational rule the case establishes: a control that is available but unmandated should be treated, for risk purposes, as absent. AISIF scopes this incident in because the Data layer extends upstream to the platforms where AI-relevant data resides.
AISIF mitigation	Governance Layer: bind platform-available controls to mandatory customer standards as a condition of production data residency; extend identity and endpoint requirements contractually and technically to contractors; inventory the datasets in the estate that feed AI workloads. Operations Layer: phishing-resistant MFA on all tenants; credential maximum lifetimes with revocation triggered by infostealer marketplace intelligence; trusted-location network policies; authentication anomaly detection before egress begins. Data Layer: classify the analytical estate as AI crown-jewel data; monitor for anomalous bulk egress. Application Layer: dedicated least-privilege, short-lived service identities for AI and analytics pipelines.
Standards activated	ISO/IEC 27001:2022 Annex A (identity and access management, supplier relationships, monitoring) NIST SP 800-207 Zero Trust Architecture; MITRE ATT&CK Valid Accounts (T1078) CSA AI Controls Matrix shared-responsibility model (AI Customer role); NIST AI RMF Govern (third-party risk)

4. The AISIF Maturity Model

The AISIF Maturity Model also provides a structured self-assessment framework for understanding the current maturity of an AI security program, and the applicable highest-priority improvements. AISIF defines four maturity levels. The boundaries are score-based, derived from the weighted assessments in the AISIF assessment tool, which is released with this document.

Table 6: AISIF Maturity Model

Level	Score	What it looks like	Typical Gaps	Priority Action
Level 1 Ad Hoc	0–29%	• No AI policy. • No defined owner. • No adversarial testing. • Security incidents handled through general IT runbooks that are not calibrated for ML failures.	• Everything.	• Start with Governance • Name a system owner. • Draft an AI use policy. Complete an AI risk assessment for your highest-risk system(s).
Level 2 Controlled	30–54%	• Basic controls exist at each layer but are not connected. • Policies exist but are not enforced. • OWASP is referenced in reviews but not embedded in the SDLC.	• Integration between layers. • Monitoring as afterthought. No feedback loop.	• Implement continuous output monitoring. • Connect monitoring findings to a governance review cycle. • Embed OWASP LLM Top 10 in code reviews.
Level 3 Integrated	55–79%	• AISIF architecture is substantively in place. • Recursive feedback loop is functioning. • NIST AI RMF Govern implemented. • Regular ATLAS-informed red team exercises.	• ISO 42001 is not yet certified. • Some data lineage gaps. Monitoring is manual rather than automated.	• Pursue ISO 42001 certification. • Automate drift detection. • Run quarterly adversarial simulation exercises.
Level 4 Adaptive	80–100%	• Board-level AI risk reporting. • ISO 42001 certified. Automated feedback loop. • Continuous adversarial simulation. • Model SLOs include security dimensions.	• Sustaining the posture as AI systems and threats evolve. • Extending to new system types, particularly agentic AI.	• Extend AISIF to cover agentic AI. • Build red team automation. • Contribute maturity evidence to regulatory reporting.

How to Perform a Self-Assessment

The fastest way to assess your current maturity level is to work through the twenty-six diagnostic questions below (instrument v0.2), ensuring responses are based on operational realities, not what is documented in policies and standards. Three rules make the assessment honest. First, answer with evidence, not intent: a control that is planned, available but unmandated, or documented but unexercised is answered No — a qualified answer is a No. Second, answer per layer and per estate: score your own infrastructure and your SaaS or third-party data estate separately, because Case Study E demonstrates these can sit at different maturity levels simultaneously. Third, report maturity per layer, not as a single aggregate: Case Study D demonstrates that high capability in one layer can coexist with Ad Hoc indicators in another. Questions marked as Level 2 floor questions (1, 2, 3, 20, and 22) are proposed gating questions: an organisation that cannot answer all five affirmatively remains at Level 1 regardless of its other answers. The instrument is also available as a formula-driven Excel workbook with a per-layer, per-estate dashboard (see the Resources appendix).

Table 7: Maturity Self-Assessment Questions (Instrument v0.2 — 26 Questions)

#	Self-Assessment Question	Maturity Linkage
Governance Layer		
1	Do you maintain a complete inventory of every AI system in use — including embedded SaaS AI features, AI agents, and MCP/tool integrations — with a named accountable owner recorded for each?	**Level 2 floor (proposed)**
2	Is there a standing AI governance body with defined membership, meeting cadence, and documented authority to accept, reject, or condition AI risk?	**Level 2 floor (proposed)**
3	Can you name the individual accountable for each AI system's security posture, and confirm they have reviewed its risk assessment in the last six months?	**Level 2 floor (proposed)**
4	Does every AI system pass a defined risk-assessment and approval gate before production, and does that gate operate at deployment cadence rather than on a periodic review calendar?	Scored
5	For every third-party AI or data platform, is there a documented shared-responsibility assignment, and are platform-available security controls — MFA, network policy, logging — made mandatory by your own standards rather than left to default configuration?	Scored
6	Are contractors and other third parties accessing your AI systems or data estate held to the same identity and endpoint standards as your employees?	Scored
Data Layer		
7	Can you trace every record in your current training dataset back to its source?	Scored

8	Are all datasets feeding AI workloads — training corpora, fine-tuning sets, feature stores, and retrieval indexes — inventoried and classified, including derived and ancillary stores such as logs, traces, and telemetry?	Scored
9	Are chat content, personal data, and secrets redacted or minimised at the point of log and telemetry ingestion, rather than persisted in plaintext?	Scored
10	Would you know within 24 hours if a data distribution anomaly occurred in your inference pipeline?	Scored
11	Would you detect anomalous bulk egress from the data estate that feeds your AI workloads before external parties told you about it?	Scored
Model Layer		
12	Has an adversarial test been performed against your model(s) using ML-specific attack scenarios from MITRE ATLAS in the last 12 months — not just a standard application penetration test?	Scored
13	Is model provenance verified — models sourced from approved registries, with an AI bill of materials maintained for production systems?	Scored
14	Are model changes — fine-tunes, system prompt revisions, and version upgrades — subject to evaluation and approval before release, with the ability to roll back?	Scored
Application Layer		
15	Are prompt injection defenses implemented at the application layer?	Scored
16	Is retrieved RAG content treated as untrusted input before it is passed to the model?	Scored
17	Are model outputs consumed by downstream systems — code, queries, commands, URLs — validated or sandboxed before execution, never executed blindly?	Scored
18	Do AI agents operate under dedicated least-privilege identities, invoke only allowlisted tools, and require human approval for irreversible or high-impact actions?	Scored
19	Are all credentials used by AI systems vaulted and short-lived, with secrets prohibited from prompts, logs, and agent-accessible storage?	Scored
Operations Layer		
20	Is every data store in the AI serving path — including analytics, observability, and telemetry stores — authenticated, network-restricted, and covered by MFA where applicable?	**Level 2 floor (proposed)**
21	Do you operate continuous external attack-surface monitoring — would you discover an exposed AI-adjacent asset before an outside researcher does?	Scored
22	Are AI interactions — prompts, outputs, tool calls, and agent decisions — logged and integrated into your security monitoring, with alerting on anomalous behaviour?	**Level 2 floor (proposed)**
23	Do credentials for AI systems and their data estate have defined maximum lifetimes, scheduled rotation, and revocation triggered by compromise intelligence?	Scored
24	Does every production AI system have a tested disablement mechanism and an AI-specific incident response playbook?	Scored

25	Are operational monitoring findings reviewed with appropriate business stakeholders, in your governance risk-management cycle, on a defined schedule?	Scored
26	Does that cycle have the authority to trigger control updates?	Scored

Maturity Progression: The Three Critical Transitions

Level 1 to Level 2

This transition requires establishing three things:

- A documented AI use policy
- A named accountability owner, and
- A completed AI risk assessment.

These governance prerequisites make systematic control implementation possible. Without them, controls will exist in isolation and have no coherent program around them. These prerequisites are organizational rather than technical and they require executive sponsorship and a decision about who owns AI risk.

Level 2 to Level 3

This is the most difficult transition in the model, and the one where most organizations have challenges. It requires implementing the recursive feedback loop, in which operational monitoring findings are reviewed by the governance function, with the authority and process to update controls in response. This requires cross-functional collaboration between security operations, data engineering, ML engineering, and governance teams who often have separate reporting lines, separate tools, capabilities and priorities, often with little or no existing process for connecting their work. Although the technical controls needed for Level 3 are not complex, organizational integration required usually is.

Level 3 to Level 4

Level 4 requires automating the feedback loop, pursuing ISO 42001 certification, and embedding AI security into strategic decision-making at the board level. This typically requires board-level sponsorship, dedicated AI security team capability, and a formal certification program. For many organizations, reaching and sustaining Level 3 is the appropriate near-term objective. Level 4 is the right target for organizations with extensive AI deployment in regulated sectors, or those subject to EU AI Act high-risk obligations.

5. AISIF Control Domains and Standards Crosswalk: ISO 27001, NIST AI RMF, and ISO 42001

AISIF introduces ten control domains aligned with its five-layer architecture. Detailed descriptions of each control domain, along with their corresponding layers, are presented in Table 8. Subsequently, relevant controls are drawn from ISO/IEC 27001:2022 Annex A, NIST AI Risk Management Framework, and ISO/IEC 42001:2023, and mapped to each control domain. Table 9 provides a consolidated overview of these crosswalk mappings, illustrating the relationships between AISIF control domains and layers and the corresponding controls within the referenced standards and frameworks.

Table 8: AISIF Control Domain Descriptions

Layer Name	Control Domain	Description
Governance	AI Risk Governance & Lifecycle Management	Establishes the organizational accountability structures, risk management cycle, and policy framework that govern the AI system from initial deployment through ongoing operation, retraining, and eventual decommissioning — ensuring that security decisions have a named owner, a defined process, and a feedback mechanism that keeps the program current as the system and its threat landscape evolve. This is the master control domain from which all other AISIF controls derive their authority and their connection to organizational risk appetite
	Shared Responsibility & Ownership	Documents the precise boundary between the organization's security obligations and those of the AI platform provider, cloud vendor, or third-party model supplier — making explicit which controls belong to each party across every layer of the AI stack. Without this mapping, accountability gaps at deployment boundaries are assumed away rather than addressed, and controls that no party owns are controls that no party implements
	Regulatory Compliance & Documentation	Maintains the documentation, evidence, and audit trail required to demonstrate that the AI system meets applicable legal and regulatory obligations — including EU AI Act risk tier requirements, sector-specific guidance such as OSFI and Health Canada AI frameworks, and the management system requirements of ISO 42001. Distinguishes compliance demonstration, which requires structured documentation of governance and controls, from security assurance, which requires those controls to function.
Data	Data Governance & Lineage	Establishes documented provenance, access control, and quality assurance for all data entering the AI system's training and inference pipelines — ensuring that every record in a training dataset is traceable to a verified source and that data transformations are logged end-to-end. Without this control, the model's trustworthiness cannot be asserted independently of the data estate that shaped it
	Training Data Integrity & Poisoning Defense	Protects the training pipeline against the deliberate or accidental introduction of corrupted, adversarially crafted, or statistically anomalous records that can shift a model's decision boundary in ways that remain invisible through standard evaluation metrics. Implements hash-based batch validation, distribution monitoring, and pre-deployment decision boundary analysis to detect poisoning before a compromised model reaches production.
Model	Threat Modeling & Attack Simulation	Systematically identifies and evaluates ML-specific adversarial threats against AI systems using the MITRE ATLAS taxonomy — covering

Layer Name	Control Domain	Description
		reconnaissance, evasion, poisoning, extraction, and model theft. Goes beyond traditional application security testing to simulate the attack techniques an adversary would realistically use against a deployed AI system's full threat surface
	Model Access Control & Authentication	Enforces least privilege access to model weights, inference endpoints, fine-tuning interfaces, and model management APIs — ensuring that only authorized identities and systems can query, modify, or retrieve model artifacts. Addresses the risk that inference APIs, often treated as lower sensitivity than data stores, provides a systematic extraction surface for model theft and membership inference attacks when inadequately authenticated.
	Secure Deployment & Supply Chain	Validates the integrity of every component introduced into the AI system during development and deployment — including pre-trained model weights, third-party datasets, ML libraries, and inference infrastructure — and applies secure development lifecycle controls to the AI-specific elements of the deployment pipeline. Recognizes that model artifacts, unlike traditional software packages, can carry backdoors that survive standard code review and antivirus scanning.
Application	Prompt Injection & Input Validation	Defends against adversarial inputs — both direct user injection and indirect injection through retrieved content such as RAG documents, tool outputs, and external data sources — by enforcing structural separation between system instructions and untrusted content. Ensures that no input pathway, including plugin calls, API responses, or retrieval-augmented context, can override system prompt authority or redirect model behavior without explicit authorization
Operations	Output Monitoring & Anomaly Detection	Continuously evaluates inference outputs against defined behavioral baselines — detecting model drift, adversarial probing patterns, systematic extraction attempts, and performance degradation that would otherwise surface only through downstream business failures or user complaints. Establishes the operational signal that feeds the AISIF recursive feedback loop, connecting real-world model behavior back to the governance layer's risk management cycle

The following considerations are critical in determining the appropriate allocation of control domains across the AISIF layers:

1. **Threat Modeling** is positioned within the Model layer rather than the Governance layer, as its primary function centers on machine learning–specific adversarial analysis. Drawing on MITRE ATLAS-informed red team methodologies, it evaluates how models may be attacked, evaded, or subjected to extraction. While its outputs inform controls across all layers, its core focus is the characterization of vulnerabilities at the model level, which constitutes its principal attack surface.

2. **Secure Deployment and Supply Chain** is also situated within the Model layer rather than the Operations layer. Its primary concern lies in ensuring the integrity of model artifacts and governing their introduction into the operational environment. As such, it functions as a pre-operational control, validating inputs prior to execution, rather than monitoring system behavior post-deployment.

3. **Shared Responsibility and Ownership** resides within the Governance layer, as it fundamentally represents an accountability-mapping exercise that must precede control design. Effective control allocation across layers depends on clearly defined ownership structures; without this foundation, the assignment and enforcement of controls lack coherence and much needed accountability.

Table 9: AISIF Crosswalk with ISO 27001:2022, NIST AI RMF and ISO 42001:2023

Control Domain	AISIF Layer	ISO 27001:2022 Annex A	NIST AI RMF	ISO 42001:2023
AI Risk Governance & Lifecycle Management	Governance	A.5.1 Security policies A.5.4 Management resp. A.5.8 Project security	Full GOVERN (1.1–1.7) Full MAP (1.1–5.2) Full MANAGE (1.1–4.2)	Cl. 4–10 Full AIMS Cl. 6.1–6.3 Planning Annex A.1–A.13
Shared Responsibility & Ownership	Governance	A.5.2 Security roles A.5.31 Legal requirements A.5.19 Supplier relations	GOVERN 1.1 Policies GOVERN 6.1 Third-party MAP 1.6 Org context	Cl. 4.3 AIMS scope Cl. 5.1 Leadership Cl. 8.6 Supplier relations
Regulatory Compliance & Documentation	Governance	A.5.31 Legal requirements A.5.32 IP rights A.5.36 Policy compliance	GOVERN 1.7 Regulatory alignment GOVERN 4.1 Risk policies MAP 1.1	Cl. 4.2 Interested parties Cl. 6.1.1 Risks Cl. 10.1 Corrective action
Data Governance & Lineage	Data	A.5.12 Classification A.5.13 Labelling A.8.10 Deletion A.5.33 Records	MAP 3.5 Provenance GOVERN 1.4 Data policy MEASURE 2.8 Bias eval	Cl. 8.3 Data management Cl. 6.1.3 Quality & lineage Annex A.6 / B.7.4
Training Data Integrity & Poisoning	Data	A.8.9 Config management A.8.29 Security testing A.8.7 Anti-malware	MAP 3.5 Training provenance MEASURE 2.5 MANAGE 1.3	Cl. 8.3.2 Training controls Cl. 8.4 Impact assessment Annex A.6.2 / B.7.5
Threat Modeling & Attack Simulation	Model	A.5.7 Threat intelligence A.8.8 Vulnerability management A.5.25 Event assessment	MAP 1.5 MAP 2.3 MEASURE 2.5 GOVERN 4.2	Cl. 6.1.2 Risk assessment Cl. 8.4 Impact assessment Annex B.6.2
Model Access Control & Authentication	Model	A.5.15 Access control A.5.16 Identity management A.5.17 Authentication A.8.2 Privileged access	GOVERN 1.2 Accountability MANAGE 4.1 Incidents MEASURE 1.1 Metrics	Cl. 7.2 Competence & access Cl. 8.5.3 Access controls Annex B.6.5 / B.8.2
Secure Deployment & Supply Chain	Model	A.5.19 Supplier relations A.8.30 Outsourced dev A.8.25 Secure dev lifecycle	GOVERN 6.2 Third-party policy MAP 5.2 Deployment context MANAGE 3.2 Change management	Cl. 8.5.4 Deployment controls Cl. 8.6.2 Third-party controls Annex B.8.3
Prompt Injection & Input Validation	Application	A.8.28 Secure coding A.8.29 Security testing A.8.27 System architecture	MANAGE 2.2 MEASURE 2.6 MAP 2.3	Cl. 8.5 Design controls Cl. 9.1 Monitoring Annex B.6.1

Control Domain	AISIF Layer	ISO 27001:2022 Annex A	NIST AI RMF	ISO 42001:2023
Output Monitoring & Anomaly Detection	Operations	A.8.16 Monitoring A.8.15 Logging A.5.25 Event assessment	MEASURE 2.7 Performance MEASURE 2.9 Fairness MANAGE 2.4 Change response	Cl. 9.1 Monitoring Cl. 9.2 Internal audit Annex B.9.1 / B.9.2

Use the crosswalk in Table 9 to design AISIF-aligned controls that are simultaneously auditable against the standards your organization is accountable to.

How the Three Standards Interrelate

A recurring question among practitioners reviewing this crosswalk is whether organizations that have achieved certification to ISO/IEC 27001:2022 must also pursue ISO/IEC 42001:2023, and how the NIST AI Risk Management Framework aligns with both. In essence, these three standards address distinct but complementary dimensions of the assurance landscape and are most effective when implemented in an integrated manner.

ISO/IEC 27001:2022 (The Foundational Security Baseline)

ISO/IEC 27001 serves as the foundational layer of security governance. Its Annex A controls establish a comprehensive framework for managing information security risks across data, systems, personnel, suppliers, and organizational processes. Within the context of AI initiatives, it provides the essential baseline upon which AI-specific controls can be meaningfully constructed. Organizations that are already certified possess an Information Security Management System (ISMS) that can function as the structural platform for integrating AISIF controls. For those not yet certified, the development of AI security capabilities and ISO 27001 implementation should ideally proceed in parallel. The 2022 revision introduced several controls of particular relevance to AI, including A.5.7 (Threat intelligence), A.8.11 (Data masking), and A.5.19 through A.5.23 (Supplier security). While indispensable, ISO 27001 alone is not sufficient for comprehensive AI security, as it does not explicitly address machine learning–specific threats, model lifecycle risks, or AI governance considerations.

NIST AI Risk Management Framework (The AI Risk Governance Cycle)

The NIST AI RMF operates at a higher level of abstraction, focusing on the governance and lifecycle management of AI-related risks. Whereas ISO 27001 emphasizes the existence and implementation of controls, the AI RMF emphasizes the systematic identification, assessment, and management of AI-specific risks. Its four core functions align closely with AISIF:

- Govern establishes accountability and oversight structures within the Governance layer.

- Map defines system context and risk exposure, informing both the Data and Model layers
- Measure evaluates the effectiveness of controls across all layers, and
- Manage ensures that risk responses are implemented and continuously refined through feedback loops between Operations and Governance.

Importantly, the AI RMF does not replace ISO 27001 controls; rather, it provides direction regarding where and why such controls should be applied. For organizations initiating an AI risk governance program, prioritizing the Govern function is essential before extending into broader framework adoption.

ISO/IEC 42001:2023 (Artificial Intelligence Management System (AIMS))

ISO/IEC 42001 represents the most AI-specific and recent addition to this triad. It defines the requirements for establishing, implementing, and maintaining an Artificial Intelligence Management System (AIMS), enabling organizations to develop, deploy, and operate AI systems in a responsible and controlled manner. Structured in alignment with the High-Level Structure common to ISO management system standards such as ISO 27001 and ISO 9001, it facilitates seamless integration into existing organizational governance frameworks rather than necessitating a parallel system. In regulatory contexts, particularly with emerging frameworks such as the EU AI Act, ISO 42001 certification is increasingly recognized as a practical mechanism for demonstrating conformity. Consequently, for organizations where regulatory alignment is a key driver, ISO 42001 should be considered a strategic priority within the broader AI security program.

> ISO 27001 establishes the security foundation, the NIST AI RMF guides risk governance, and ISO 42001 formalizes AI-specific management practices. Together, they form a cohesive and mutually reinforcing framework for robust AI assurance.

6. Building Your AISIF-Aligned Program

This section translates the framework into a structured implementation sequence, wherein each step is organized according to its dependencies, with subsequent steps building upon the outputs of preceding ones. Practitioners are not required to complete all seven steps to derive value from AISIF, as each step independently produces actionable outputs that contribute to the overall program.

Table 10: AISIF's Implementation Sequence

	Seven-Step AISIF Program Implementation Sequence	Output
1	• Complete a governance foundation check. • Name the AI system(s) owner(s). • Draft or validate the AI use policy and confirm it covers the system being assessed.	• This takes one meeting if you have the right people in the room
2	• Run the AWS Scoping Matrix exercise for every AI system in your portfolio (forty-five minutes per system)	• A documented accountability map showing which controls belong to your organization and which belongs to your provider.
3	• Complete the AISIF self-assessment for each high-risk AI system (use the Excel tool released with this document).	• Output is a layer score, an overall AISIF score, a maturity level, and a prioritized gap list.
4	• Conduct a MITRE ATLAS-informed red team exercise against your highest-risk AI system(s), ensuring it is scoped explicitly to ML-specific attack vectors (evasion, poisoning, extraction, and model inversion)	• Output is a documented adversarial threat model.
5	• Implement the AISIF feedback loop. • Define the schedule on which operational monitoring findings are reviewed in your governance risk management cycle. • Define the authority and process for triggering control updates. • Document the cycle.	• Output is a functioning Operations-to-Governance feedback mechanism.
6	• Map your AISIF controls to the standards crosswalk in Section 5. • Identify which ISO 27001 controls are already in place, which NIST AI RMF subcategories are addressed, and which ISO 42001 clauses require new activity.	• Output: a gap analysis against three auditable standards.
7	• Develop a 12-month roadmap from your current maturity level to your target maturity level, using the progression guidance in Section 4. • Prioritize the transitions described there.	• Present the roadmap to the system owner for sponsorship.

Prioritization Guide: Where to Start Based on Your Biggest Gap

If the seven-step sequence feels too broad as a starting point, practitioners can use this guide to identify the single highest-leverage action based on the failure pattern most relevant to their program.

Table 11: Prioritization Guidelines

If your biggest gap is...	Your first AISIF action is...	Relevant layer and framework
You cannot name a system owner or produce a risk assessment	Establish Governance layer basics: AI policy, named owner, risk assessment	Layer 1 — Governance NIST AI RMF Govern 1.1 ISO 42001 Cl. 5.3
Your training data provenance is undocumented or ungoverned	Implement data lineage tracking and access controls on training data stores	Layer 2 — Data Databricks AI Security ISO 27001 A.5.12 / A.5.15
Your red team exercises do not cover ML-specific attack vectors	Run a MITRE ATLAS-scoped adversarial test covering evasion, poisoning, and extraction	Layer 3 — Model MITRE ATLAS NIST AI RMF MEASURE 2.6
Prompt injection controls are not implemented in your LLM applications	Implement OWASP LLM01 input sanitisation and instruction hierarchy enforcement	Layer 4 — Application OWASP LLM Top 10 ISO 27001 A.8.28
You have no continuous monitoring of model outputs post-deployment	Deploy inference logging with anomaly alerting and establish a review schedule	Layer 5 — Operations ISO 27001 A.8.15-16 NIST AI RMF MANAGE 2.4
You cannot demonstrate AI compliance to auditors or regulators	Complete the standards crosswalk and initiate ISO 42001 gap assessment	Governance + all layers ISO 42001 / ISO 27001 NIST AI RMF Govern 1.7

7. Quick Reference Cards

Table 11: AISIF in One Page

AI Security Integration Framework (AISIF) — Five-Layer Architecture		
Layer 1 — Governance	Risk management, Policy, Accountability, Regulatory compliance	*NIST AI RMF Govern · ISO 42001 · EU AI Act*
Layer 2 — Data	Lineage, Provenance, Access control, Pipeline integrity, Masking	*Databricks AI Security · ISO 27001 A.8–A.9*
Layer 3 — Model	Adversarial robustness, Provenance, Supply chain, API access	*MITRE ATLAS · Google SAIF · NIST AI RMF Map/Measure*
Layer 4 — Application	Prompt injection defense, Output filtering, Plugin security, RAG sanitization	*OWASP LLM Top 10 · Google SAIF · AWS Scoping Matrix*
Layer 5 — Operations	Monitoring, Drift detection, Incident response, Feedback loop	*Databricks · NIST AI RMF Manage · ISO 27001 A.8.15–16*
Recursive Feedback Loop	Operations findings, Governance review, Control updates, All layers	*The defining feature that distinguishes AISIF from a framework catalogue*

The 10-Point Pre-Deployment AI Security Checklist

Use this checklist before deploying any AI system, or as a rapid health check on a system already in production. It is not a substitute for the full AISIF assessment but can be a fast filter for the highest-priority gaps.

- ☐ Governance: A named owner with decision authority has been assigned and has reviewed the system's risk assessment
- ☐ **Governance:** A documented AI use policy covers this system and defines both intended and out-of-scope use cases
- ☐ **Data:** Training data provenance is documented and origin, collection method, and any known biases are on record
- ☐ **Data:** Access controls on training datasets and feature stores are enforced at the system level, not just in policy
- ☐ **Model:** An adversarial test using MITRE ATLAS ML-specific tactics has been completed in the last 12 months
- ☐ **Model:** Model artifacts (weights, checkpoints) are under version control and inference endpoints require authentication
- ☐ **Application:** Prompt injection defenses are implemented, including input sanitization and instruction hierarchy enforcement are in place

☐	**Application:** Output filtering is applied before responses reach users, including PII detection and content moderation
☐	**Operations:** Continuous inference monitoring is live with anomaly alerting configured and a defined response procedure
☐	**Operations:** A defined schedule exists for operational monitoring findings to reach the governance risk management cycle

Closing: Integration Is the Work

The frameworks referenced in this document are widely accessible. The NIST AI Risk Management Framework is publicly available, OWASP LLM Top 10 resources are open and freely distributed, MITRE ATLAS is comprehensively documented, and the EU AI Act constitutes binding legislation. The principal challenge in achieving a mature AI security program is therefore not access to frameworks, but their effective integration.

In practice, many security programs engage with these frameworks in a fragmented and episodic manner. For example, OWASP guidance may be consulted during architecture reviews, NIST AI RMF principles referenced in governance discussions, and MITRE ATLAS exercises conducted intermittently as resources permit. What is often lacking is a cohesive programmatic approach that systematically incorporates all relevant frameworks, supported by clearly defined ownership structures and sustained operational processes that extend beyond initial implementation.

AISIF addresses this gap by providing a comprehensive integration model. Its five-layer architecture establishes a structured environment in which each framework is assigned a clearly defined role. The embedded recursive feedback loop ensures that the architecture remains adaptive to the evolving characteristics of the AI systems it supports. The maturity model offers a mechanism for self-assessment and continuous progression, while the standards crosswalk enhances auditability and alignment with established benchmarks. Complementary artefacts, including the assessment tool and this document, operationalize these components within a practical implementation context.

Organizations that succeed in developing robust AI security programs will not be distinguished by the breadth of frameworks they reference, but by their ability to integrate these frameworks into a coherent and functional operating model, underpinned by clear accountability structures. AISIF is designed to enable and support this level of organizational capability.

> In the era of artificial intelligence, security can no longer be confined to the model itself. Rather, it encompasses the entire ecosystem that enables and sustains the model, including the data that informs its training, the pipelines that operationalize its deployment, the controls that govern its use, and the individuals who remain ultimately accountable for its outcomes.

Appendix: Framework and Standards References

The following references support the framework analysis in this document. All are publicly available.

MITRE

MITRE. (n.d.). *Adversarial threat landscape for artificial intelligence systems (ATLAS).* https://atlas.mitre.org

OWASP

OWASP. (2023). *Top 10 for large language model applications (Version 1.1).* https://owasp.org/www-project-top-10-for-large-language-model-applications

Google

Google. (n.d.). *Secure AI framework (SAIF).* https://safety.google/cybersecurity-advancements/saif

Amazon Web Services

Amazon Web Services. (n.d.). *Generative AI security scoping matrix.* https://aws.amazon.com/blogs/security

National Institute of Standards and Technology

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1).* https://airc.nist.gov/airmf

European Union

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act).* https://artificialintelligenceact.eu

International Organization for Standardization

International Organization for Standardization. (2022). *ISO/IEC 27001:2022 information security, cybersecurity and privacy protection—Information security management systems—Requirements.* https://www.iso.org/standard/27001

International Organization for Standardization

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 artificial intelligence—Management system.* https://www.iso.org/standard/81230.html

Databricks

Databricks. (n.d.). *Databricks AI security framework.* https://docs.databricks.com

Appendix B: AISIF New AI System Assessment Instrument (v0.2)

Where the maturity instrument in Section 4 assesses the programme, this instrument gates a single AI system before production. It asks a different question: may this system go live? The two instruments interlock through the governance gate — maturity question 4 asks whether a pre-production gate exists; this instrument is that gate.

The workflow has four steps. First, a system profile of six factors (data sensitivity, user exposure, autonomy, decision impact, third-party dependency, integration depth) computes an inherent risk tier from Low to Critical. Second, the thirty control checks below are answered with evidence — Yes, No, Planned, or N-A, where Planned deliberately does not count toward approval, because a gate must distinguish implemented from intended. Third, a provisional recommendation is computed: any Mandatory control not answered Yes (and not genuinely N-A) blocks deployment; otherwise approval requires at least 85% of applicable checks implemented, raised to 90% for High or Critical inherent-risk systems. Fourth, the governance body records its decision, conditions, approver, and re-assessment trigger — the recommendation informs the decision; it does not replace it. Criticality designations and thresholds are provisional pending practitioner validation.

Table B-1: New AI System Assessment — 30 Control Checks (17 Mandatory)

#	Control Check (for this system)	Criticality
Governance Layer		
1	Is this system registered in the AI system inventory with a named accountable business owner and a named technical owner?	**Mandatory**
2	Has a documented risk assessment for this system been completed and approved through the AI governance gate before production use?	**Mandatory**
3	For vendor or embedded SaaS AI: has third-party due diligence been completed — data handling, training-on-your-data terms, model provenance, sub-processors — and accepted?	**Mandatory**
4	Is a shared-responsibility assignment documented for every platform this system depends on, with platform-available controls (MFA, network policy, logging) made mandatory rather than left to defaults?	**Mandatory**
5	Are the intended use, user population, and explicitly prohibited uses of this system documented and communicated to users?	Recommended
6	Have privacy, legal, and regulatory reviews applicable to this system been completed?	Recommended
Data Layer		
7	Are all data sources this system ingests, retrieves, or trains on classified, and is that classification approved for this specific use?	**Mandatory**
8	Is regulated or restricted data excluded from this system unless explicitly approved with compensating controls?	**Mandatory**

9	Is input inspection (DLP) applied to prompts and uploads before they reach external models, with block or tokenise on match?	Recommended
10	Are chat content, personal data, and secrets redacted or minimised at log and telemetry ingestion for this system, rather than persisted in plaintext?	**Mandatory**
11	Do data residency and retention for prompts, outputs, and provider-side storage comply with organisational and regulatory requirements?	Recommended
12	For RAG/grounding: are retrieval sources curated and classified, with retrieval permission-trimmed to the requesting user's entitlements?	Recommended
Model Layer		
13	Is model provenance verified — sourced from an approved registry — with an AI bill of materials recorded for this system?	Recommended
14	Has adversarial testing informed by MITRE ATLAS scenarios been performed against this system before go-live — not only a standard application penetration test?	Recommended
15	Are model changes for this system — fine-tunes, system prompt revisions, version upgrades — subject to evaluation and approval before release, with rollback available?	**Mandatory**
16	For high-stakes outputs: is grounding with citations configured, with defined confidence and refusal behaviour?	Recommended
17	Is it contractually and technically confirmed that enterprise inputs to this system are not used for provider model training unless explicitly approved?	**Mandatory**
Application Layer		
18	Are prompt injection defenses implemented at the application layer for this system?	**Mandatory**
19	Is retrieved or referenced content (RAG documents, emails, web pages, files) treated as untrusted input before reaching the model?	**Mandatory**
20	Are outputs consumed by downstream systems — code, queries, commands, URLs — validated or sandboxed before execution?	**Mandatory**
21	If agentic: does the system run under a dedicated least-privilege identity, invoke only allowlisted tools, and require human approval for irreversible or high-impact actions? (Answer N-A if not agentic.)	**Mandatory**
22	Are all credentials used by this system vaulted and short-lived, with secrets prohibited from prompts, logs, and agent-accessible storage?	**Mandatory**
23	Does the system enforce the requesting user's existing entitlements — no privilege escalation through the AI layer?	Recommended
Operations Layer		
24	Is every datastore in this system's serving path — including analytics, observability, and telemetry stores — authenticated, network-restricted, and covered by MFA where applicable?	**Mandatory**
25	Are this system's interactions — prompts, outputs, tool calls, agent decisions — logged and integrated into security monitoring, with alerting on anomalous behaviour?	**Mandatory**
26	Do the system's credentials and service identities have defined maximum lifetimes, scheduled rotation, and revocation triggers?	Recommended

27	Has an external attack-surface review of the system's internet-facing components been performed before go-live, with continuous exposure monitoring in place?	Recommended
28	Does the system have a tested disablement mechanism (kill switch) and an AI-specific incident response playbook naming this system?	**Mandatory**
29	Are operational support ownership, escalation paths, and a runbook defined for this system?	Recommended
30	Is there a defined schedule for this system's monitoring findings to reach the governance cycle, with authority to trigger control updates?	Recommended

Appendix C: AISIF Resources and Assessment Tools — How to Use Them

All AISIF materials are released as open, versioned drafts. The assessment instruments carry provisional scoring pending validation against practitioner data — organisations using them are invited to share anonymised results, which is the validation path for the thresholds. The intended operating rhythm: run the maturity self-assessment annually at programme level; run the system assessment for every new AI system, every time, before production; and let the recursive feedback loop route what both instruments find into control updates.

Table C-1: AISIF Resource Catalogue

Resource	What it is	How to use it
AISIF Practitioner White Paper (this document)	The full framework: five-layer architecture, recursive feedback loop, control domains, standards crosswalks, maturity model, and case studies A–E.	Read Sections 1–2 for the argument, Section 5 for the crosswalk when mapping to ISO 27001, NIST AI RMF, or ISO 42001, and Section 6 to sequence your programme build.
Case Study Series (A–E)	Three illustrative composites (A: prompt injection; B: data poisoning; C: model inversion) and two empirical analyses of disclosed incidents (D: DeepSeek infrastructure exposure; E: Snowflake identity campaign), each in the structured card format with standalone documents available for D and E.	Use in tabletop exercises and awareness briefings; use D and E when making the business case that AI security budgets must fund conventional controls, not only AI-specific tooling.
Maturity Self-Assessment Instrument (Word + Excel workbook)	The 26-question instrument in Section 4, also delivered as a standalone document and a formula-driven Excel workbook with an Introduction/metadata tab, one tab per layer, and a dashboard computing per-layer and per-estate results, Level 2 floor status, feedback-loop status, and provisional maturity levels.	Run annually and after material change. Interview control owners with evidence in hand; complete both estate columns; report the per-layer profile, not an average. Your weakest layer is your programme's real level.
New AI System Assessment Workbook (Excel)	The per-system deployment gate in Appendix B as a workbook: system identification, a System Profile tab computing the inherent risk tier, the 30 control checks with evidence columns, and a dashboard producing the provisional recommendation and governance sign-off block.	Run for every new AI system before production and re-run on material change. Route the completed dashboard to the AI governance body as the decision record. A blocked recommendation is the instrument working.
Top-25 AI Guardrails Set	A curated control set across six domains (governance, data protection, identity and access, input/output runtime controls,	Use as the control skeleton for an enterprise AI guardrails standard; map each guardrail to your control IDs and

	agentic and tool controls, monitoring and resilience), traceable to the CSA AI Controls Matrix, OWASP GenAI, NIST AI 600-1, and MITRE ATLAS.	cite the source frameworks in an appendix for auditability.
Conference Presentations (BSides Edmonton 2026; SecTor 2026)	A 24-slide deck presenting the framework, the empirical case studies, and both assessment instruments, with speaker notes.	Reuse internally for executive and team briefings; the three-actions slide (inventory, maturity self-assessment, gate the next system) is a ready-made 30-60-90 day starting plan.
AISIF Website (forthcoming)	A central download point for all of the above, with versioned releases.	Check for the latest instrument versions before each assessment cycle; thresholds and criticality designations will be updated as practitioner validation data accumulates.