What REALLY Matters in AI Security: A Risk-First Mental Model

Executive Summary

AI security has become a budget priority across enterprises, yet organizations struggle to separate signal from noise. This paper provides a risk-first mental model that maps AI security investment to actual loss impact, grounded in NIST CSF 2.0, OWASP guidance, and current threat intelligence.

The core thesis: AI primarily amplifies existing cyber threats rather than creating fundamentally new loss channels. Effective AI security requires three sequential layers:

1. **Cyber residual risk posture** – Harden the foundational environment
2. **Fundamental AI/app security** – Address AI-native technical risks
3. **Technical and operational governance** – Run AI as a managed program

TL;DR: Don't talk AI security until you've stabilized the cyber residual risk floor that AI amplifies, secured AI-native architectures against OWASP/agentic failure modes, and wrapped it all in governance mapped to NIST AI RMF and CSF 2.0. ***Except that NHI IAM is needed ASAP!*** *ALSO, the Cloud Security Alliance's AI Controls Matrix (AICM) provides the control-objective catalog underneath this model: it takes the same risk-first priorities and expresses them as concrete, auditable AI and cloud controls across the lifecycle*

NOTE – In addition to the **AICM** that gives us a comprehensive, lifecycle-wide catalog of AI and cloud control objectives, mapped to NIST CSF 2.0, NIST AI 600-1, ISO 42001. Enterprise teams building *agentic* systems need a more opinionated, implementation-ready checklist. That is where the **Top 20 AI Agent Security Controls** come in: they are a condensed, operational subset of AICM and OWASP LLM guidance, focused specifically on agent identity, prompts and tools, runtime isolation, memory, oversight, supply chain, and compliance

ANSWER - What "Really" Matters in AI Security

AI has changed the speed, scale, and accessibility of cyberattacks—but it has not changed where most of the money is actually lost. Today, the majority of AI-boosted incidents still monetize through the same old paths: phishing and business email compromise, credential abuse, vulnerability and misconfiguration exploitation, and ransomware/data theft. The practical question is not "What is the cleverest new AI attack?" but **"Which controls most reduce AI-driven loss?"**

A risk-first view of current incident and breach-cost data suggests three layers of control do the heavy lifting:

- **Layer 1 – Cyber residual risk floor (~60–70% of avoidable AI-driven loss):** Hardening identity, patching, misconfigurations, backups, and email/endpoint security still delivers the majority of achievable AI risk reduction, because it denies AI-assisted attackers the initial access and blast-radius they depend on.
- **Layer 2 – AI / app security (~15–25%):** AI-specific controls (prompt-injection defenses, safe tool integration, AI-generated code review, agentic guardrails) provide a meaningful second step, reducing incremental loss from AI as a new attack surface and execution engine.
- **Layer 3 – Governance, shadow AI, and NHI IAM (~10–20%):** Program-level governance—AI inventory, risk acceptance, non-human identity management, cross-border and third-party controls—moves a smaller slice of everyday loss, but disproportionately reduces tail-risk: regulatory, legal, and reputational "AI disasters."

These figures are directional bands based on current incident and cost data, not precise actuarial estimates. Put simply: **don't talk AI security until you've stabilized the cyber residual risk floor.** Once the environment is hardened, secure how AI systems and agentic apps are built and integrated. Only then does governance—shadow AI control, NHI IAM, cross-border policy—deliver its full value without becoming security theater.

Introduction: The AI Security Investment Paradox

AI for security and security for AI is now the top cybersecurity budget priority, cited by 36% of executives—ahead of cloud security (34%), network security (28%), and data protection (26%)[22][28]. Yet only 6% of organizations consider themselves "very capable" to withstand AI-enabled attacks across all vulnerabilities[28], and 83% lack comprehensive AI security visibility[3].

This paradox stems from a fundamental misalignment: **organizations are investing in AI security before understanding what AI actually threatens.**

Most near-term AI risk is **loss amplification on existing cyber weaknesses**, not science-fiction model takeovers[1][10]. AI turns commodity threats—phishing, credential abuse, misconfiguration exploitation, vulnerability chaining—into always-on, high-volume attack pipelines[1][8][10][14]. The economic impact is measured in billions:

- Business Email Compromise (BEC) caused $2.7–$6.3 billion in U.S. losses in 2024, with AI-generated BEC emails now comprising 40% of attacks[36][40][44]
- Average phishing-related breach costs $4.88 million, with AI weaponization driving a 1,265% surge in phishing volume[36]
- AI-powered cyberattacks contributed to $10.5 trillion in projected losses in 2025, with around one-half of attacked small businesses closing within six months[21][27]
- Shadow AI data leakage costs organizations an average of $650,000 per breach[29]

The financial toll is staggering not because AI invents new attack classes, but because it **industrializes exploitation of unpatched, misconfigured, and identity-weak environments**[1][10][21].

This paper provides a three-layer mental model that aligns AI security investment with actual loss drivers and maps cleanly to NIST CSF 2.0, OWASP guidance, and emerging AI risk frameworks.

*This paper deliberately treats **AICM** as an implementation backbone rather than the starting point. Our three-layer model answers "where does AI actually move loss and in what order should we invest?"; AICM then supplies the detailed control objectives, ownership tags, and standard mappings (NIST CSF 2.0, NIST AI 600-1, ISO 42001, AI Act) so organizations can turn that risk-first mental model into concrete policies, technical safeguards, and vendor requirements without inventing their own control library from scratch.*

What Does the Research Say?

AI Primarily Amplifies Existing Threats

The UK National Cyber Security Centre's 2024 assessment concluded that AI's near-term impact is to **lower barriers and increase volume** for established attack types rather than enabling fundamentally new threats[1]:

- AI-enhanced phishing and social engineering: better localization, more convincing lures, higher volume at lower cost
- Faster reconnaissance and exploitation: automated vulnerability discovery, exploit generation, and credential stuffing at scale
- Commodity vulnerability chaining: AI assists in identifying and chaining known CVEs into novel attack paths

Multiple 2025-2026 threat intelligence reports confirm this pattern[8][10][14][25]. Google's March 2026 analysis documented state-sponsored and cybercriminal groups using AI to increase attack precision and scale, particularly in phishing campaigns and malware development[8].

The Three Primary Loss Drivers

Analysis of recent breach data, threat reports, and industry surveys reveals three dominant AI-related loss mechanisms:

1. Identity and credential compromise amplification

AI dramatically accelerates credential-based attacks:

- Phishing emails that bypass traditional detection (1,265% volume increase)[36]
- Deepfake-enabled BEC and social engineering (53% of accounting professionals experienced deepfake attacks in 2024)[44]
- Automated credential stuffing and password spraying against exposed identity systems[1][10]

Average BEC wire transfer requests climbed to $80,000–$129,000 in 2024-2025, with individual attacks often unrecoverable[36][40].

2. Misconfiguration and vulnerability exploitation at scale

AI-assisted reconnaissance and exploitation reduces time-to-compromise:

- Automated discovery of cloud misconfigurations (S3 buckets, exposed databases, overpermissioned IAM roles)[1][10][11]
- AI-generated exploit code for known vulnerabilities, including KEV catalog items[1][4][10]
- Supply chain attacks leveraging poisoned models and datasets in public repositories[24][30]

Organizations using extensive security AI and automation saved an average of $2.2 million in breach costs compared to those without such technologies, highlighting the cost of inadequate foundational controls[39].

3. Unmanaged AI adoption (Shadow AI)

Employee use of unauthorized AI tools creates immediate data exposure:

- 20% of UK companies experienced data leakage from employees using GenAI without oversight[23]
- 59% of organizations adopt AI faster than they can secure it[26]
- Shadow AI risks include intellectual property exposure, compliance violations, and loss of data sovereignty[23][29]

IBM's 2025 Cost of Data Breach Report found Shadow AI exposure cost organizations an average of $650,000 per breach due to lack of AI governance frameworks[29].

Emerging AI-Native Risks

While amplification dominates current losses, AI-specific technical vulnerabilities are materializing:

- **Prompt injection and data exfiltration:** Indirect prompt injection attacks against Microsoft Copilot, Google Bard/Gemini, Claude, Notion AI, and others successfully exfiltrated sensitive data in 2024-2026[37][41][45]
- **Model and data poisoning:** Researchers demonstrated supply chain attacks embedding command-and-control servers in public models and publishing poisoned datasets to PyPI[24][30]
- **Excessive agency and tool misuse:** AI agents granted broad access to email, customer records, and APIs enable single-click data exfiltration via injected prompts[37][38][45]

These OWASP Top 10 for LLM risks are real but remain secondary to identity, misconfiguration, and shadow AI losses in aggregate economic impact[7][13][17].

AICM and the "Top 20" AI Controls

The Cloud Security Alliance's **AI Controls Matrix (AICM)** gives us a comprehensive, lifecycle-wide catalog of AI and cloud control objectives, mapped to NIST CSF 2.0, NIST AI 600-1, ISO 42001, and the EU AI Act. It ensures we are not inventing our own control library from scratch.

At the same time, enterprise teams building *agentic* systems need a more opinionated, implementation-ready checklist. That is where the **Top 20 AI Agent Security Controls** come in: they are a condensed, operational subset of AICM and OWASP LLM guidance, focused specifically on agent identity, prompts and tools, runtime isolation, memory, oversight, supply chain, and compliance.

Conceptually, the relationship is simple:

- The **three-layer model** tells us *where* AI risk actually moves loss (Layer 1 residual cyber risk, Layer 2 AI/app security, Layer 3 governance and NHI IAM) and in what sequence to invest.
- **AICM** provides the **full control-objective mesh** underneath that model across the AI lifecycle and shared responsibility roles.
- The **Top 20 AI Agent Controls** are a **Layer-2/Layer-3 "fast path"** for agentic workloads: they select and group the AICM and OWASP-aligned controls that most directly reduce agent-driven loss and regulatory tail-risk in practice.

In other words, AICM is the *complete* map; the Top 20 are the **shortest path through that map** for teams deploying agentic systems in 2026, and our three-layer model explains why these particular controls deserve to be first in line.

Appendix D provides a simple lookup table from each Top-20 agent control to its primary layer and the most relevant AICM domain(s), so teams can go straight from this model to specific AICM rows and evidence expectations.

The Three-Layer Mental Model

Layer 1: Cyber Residual Risk Posture (Foundational Hygiene)

Purpose: Establish the secure environment baseline that prevents AI from amplifying existing weaknesses.

Core question: Is our environment hardened enough that AI-specific controls aren't theater?
This layer addresses the attack surface AI-assisted adversaries will opportunistically weaponize. It aligns with NIST CSF 2.0 core functions (Identify, Protect, Detect, Respond, Recover) and emphasizes: ***NOTE – an effort to get NHI IAM technical guidance in place asap is needed!***

Control Domain	Key Activities
Identity & Access Management	Multi-factor authentication (MFA) enforcement, least privilege access, privileged access management (PAM), credential rotation, password policy hardening
Vulnerability & Patch Management	CISA KEV catalog remediation, automated vulnerability scanning, risk-based patch prioritization, secure baseline configurations
Cloud Security Posture Management	Misconfiguration detection (exposed storage, overpermissioned roles), CSPM continuous monitoring, infrastructure-as-code security scanning
Email & Endpoint Protection	Anti-phishing controls, endpoint detection and response (EDR), email authentication (SPF/DKIM/DMARC), security awareness training
Logging, Backup & Recovery	Centralized logging with tamper-proofing, immutable backups, tested recovery procedures, incident response playbooks

Table 1: Layer 1 control domains and activities

Most AICM controls that are tagged as "cloud-only" or "AI+cloud" fall into this Layer 1 foundation: identity and access, vulnerability and misconfiguration management, logging, and recovery. In practice, these AICM items should be treated as hygiene baselines and cross-checks on your existing CSF/CCS posture, not as "AI features" – until this floor is stable, AI-specific control shopping is largely security theater.

Why this matters for AI security: Every major AI-enabled breach in 2024-2026 began with exploitation of identity weaknesses, unpatched vulnerabilities, or cloud misconfigurations[1][10][21][36]. Organizations with strong Layer 1 posture reduce breach costs by an average of $2.2 million[39].
Alignment: CIS Controls IG1, NIST CSF 2.0 core functions, Cloud Cyber Shield (CCS) framework, ENISA threat landscape recommendations[1][10][11][12].

Layer 2: Fundamental AI / App Security (Agentic Patterns)

Purpose: Address AI-native technical risks in data, models, context assembly, tools, and orchestration.
Core question: Are our AI systems and agentic applications built and integrated securely?
This layer focuses on vulnerabilities unique to LLM-based systems and agentic AI workflows. It directly implements OWASP Top 10 for LLM Applications and emerging agentic AI security guidance:
At the control-objective level, Layer 2 is where AICM's AI-specific domains—Model Security, AI Supply Chain Management, Data Security & Privacy, and Application & Tool Security—live, providing concrete safeguards for the risk categories described here.

Risk Category	Mitigation Strategies
Prompt Injection	Input sanitization, output validation, context isolation, user permission boundaries, monitoring for injection patterns
Insecure Output Handling	Escape LLM outputs before rendering, avoid direct code/command execution, sandboxed tool invocation
Training Data & Model Poisoning	Sanitize public training data, validate model behavior, supply chain verification for models and datasets

Model Denial of Service	Rate limiting, resource quotas, input length restrictions, cost monitoring and alerting
Supply Chain Vulnerabilities	Model provenance verification, dependency scanning, private model registries, artifact signing
Insecure Plugins & Tools	Least privilege for tool access, input validation for tool parameters, audit logging of tool invocations
Excessive Agency	Human-in-the-loop for high-impact decisions, scope limitations, time-bound access, autonomous action logging
Sensitive Data Disclosure	Data Loss Prevention (DLP) integration, prompt filtering for PII/secrets, RAG data access controls

Table 2: Layer 2 AI-native risk categories and mitigations

These risk categories map cleanly onto AICM: prompt injection, insecure output handling, and sensitive data disclosure align with AICM's model and application controls; model and data poisoning, insecure plugins/tools, and supply-chain vulnerabilities align with AI supply-chain and artifact integrity controls; and excessive agency maps to AICM's tagging of control ownership across model providers, orchestrators, application teams, and AI customers under its Shared Security Responsibility Model.

Layer 2 control mapping (Top 20 → technical implementation)
At Layer 2, "good enough" means the following Top 20 controls are implemented for each production agent and AI workflow:

- **Identity & Access (1–3):** Dedicated non-human identities for every agent with strong auth (#1), least-privilege tool access (#2), and signed, authenticated inter-agent calls (#3).
- **Input / Output (4–6):** Prompt-injection defenses on all inputs (#4), schema and safety validation on all outputs (#5), and PII/secrets detection and scrubbing before prompts, memory, or third-party calls (#6).
- **Ops & Isolation (7–10):** Explicit action allowlists and guardrails (#7), rate limiting and throttling (#8), sandboxed and isolated runtimes (#9), and immutable audit logs of every tool call (#10).
- **Memory & State (11–12):** Encrypted, access-scoped vector/RAG stores (#11) and session/state integrity checks to prevent poisoning and cross-session leakage (#12).
- **Supply Chain & Reliability (16–18, technical side):** Vetted and pinned models (#16), hallucination and reliability controls for high-stakes outputs (#17), and monitoring for model behavioral drift (#18).

Implementation detail and AWS mappings for each of these controls are captured in **Appendix A.**

Real-world impact: Prompt injection vulnerabilities enabled single-click data exfiltration from Microsoft Copilot, Google Bard, Claude Cowork, and other enterprise AI tools in 2024-2026[37][41][45]. Model supply chain attacks allowed researchers to compromise organizations by uploading poisoned models to public registries[24][30].
Alignment: OWASP Top 10 for LLM Applications, SANS "Protect AI" domain, NIST AI RMF "Manage" function, AI code generation security guidance[4][7][9][13][17].

Layer 3: Technical and Operational Governance (Risk, Policy, Assurance)

Purpose: Run AI as a managed program with accountability, regulatory compliance, and lifecycle oversight.
Core question: Are we governing AI with appropriate risk acceptance, policy enforcement, and continuous assurance?

This layer provides the connective tissue between technical controls (Layers 1 and 2) and organizational risk management. It maps directly to NIST AI RMF governance functions and CSF 2.0 "Govern" category: It focuses on a small set of governance domains that every AI program must cover.
AICM contributes here by expressing governance as explicit, testable control objectives: inventory and classification controls, risk assessment and acceptance controls, transparency and accountability controls, and compliance/alignment controls tagged by lifecycle stage and responsible party. The same AICM matrix that describes model and data safeguards also tells you who (CSP, model provider, orchestrated service, application owner, or AI customer) is expected to implement which governance activities at which point in the AI lifecycle.

Governance Domain	**Key activities** (what "good" looks like)
AI Inventory & Classification	Maintain a registry of AI systems and use cases, classify by risk tier (e.g., high/medium/low), and assign clear ownership and accountability.
Risk Assessment & Acceptance	Perform AI-specific risk assessments (e.g., data misuse, model abuse, safety/ethics risks), document risk acceptance decisions, and escalate high-risk use cases for centralized review.
Policy & Standards	Define acceptable-use policies, AI design and secure development lifecycle (SDL) requirements, and data-handling standards for training, fine-tuning, RAG, and logging.
Regulatory & Compliance Alignment	Map to sector-specific regulations (GDPR, HIPAA, financial services), and internal policies, implement audit trails, prepare for emerging AI regulation and customer/partner due diligence.
Continuous Monitoring & Assurance	Monitor key AI behaviors (e.g., anomalous outputs, model drift, incident patterns, Shadow AI), run periodic security/governance reviews, and keep AI-specific incident response playbooks current.
Identity Governance for AI Agents (NHI IAM)	Treat AI agents and services as first-class non-human identities: enforce least privilege, require clear human ownership, use time-bound access, and maintain end-to-end audit trails of actions.

Table 3: Layer 3 governance domains and activities

Layer 3 control mapping (Top 20 → governance implementation)
At Layer 3, a credible AI governance program ensures:

- **Oversight (13–15):** Human-in-the-loop checkpoints for irreversible actions (#13), tested kill switches at agent/session/global levels (#14), and rollback and action reversibility via compensating transactions and logs (#15).
- **Compliance (19–20):** AI-specific incident response categories and notification paths (#19), plus transparency and plain-language explainability when users interact with AI and when high-risk decisions are made (#20).
- **Supply Chain & Drift (16–18, governance side):** Governance over LLM provider vetting and data-use clauses (#16), policies for hallucination controls in critical workflows (#17), and oversight of silent model updates and statistical drift (#18).
- **NHI IAM & AI-SPM (cross-cutting):** All AI agents and services are governed as non-human identities with lifecycle, least privilege, and audit trails, continuously monitored in production by AI-SPM for inventory, control coverage, and regulatory alignment.

Concrete governance methods, NHI IAM patterns, AWS controls, and AI-SPM capabilities for these controls are documented in **Appendix B**

In practice, organizations can take the simple Layer 3 questions in this paper—"Do we have an inventory?", "How do we accept AI risk?", "What are our NHI IAM standards?"—and bind them directly to AICM control IDs. Using AICM's Shared Security Responsibility Model as the RACI spine ensures that each inventory, policy, and identity control has a clear owner across CSPs, model providers, internal platform teams, and consuming business units, avoiding the familiar "everyone and no one owns AI" problem

Anchoring Layers 2 and 3 in concrete AI agent controls
To keep this model actionable, we anchor Layers 2 and 3 to a specific control catalogue: the *Top 20 AI Agent Security Controls* (Identity & Access, Input/Output, Ops & Isolation, Memory & State, Oversight, Supply Chain, Compliance). These controls become the explicit, testable requirements for how we design, build, and govern agentic AI systems.

- **Layer 2 – AI / Application Security** implements the technical half of the catalogue: Identity & Access (1–3), Input/Output (4–6), Ops & Isolation (7–10), Memory & State (11–12), and the technical portions of Supply Chain & Reliability (16–18).
- **Layer 3 – Technical & Operational Governance** implements the oversight and assurance half: Oversight (13–15), Compliance (19–20), and the governance aspects of Supply Chain and Drift (16–18), including NHI IAM, AI inventory, risk tiering, and AI-SPM.

We measure Layer 2 and 3 maturity by asking, for each agent and AI use case, which of these 20 controls are implemented, monitored, and evidenced—and which are not.

How we implement Layers 2 and 3 (methods, tools, and platforms)
For each AI agent control, we define: a **method** (what the control must achieve), a set of **generic tool patterns** (e.g., API gateways and policy engines for action guardrails, containers and network segmentation for sandboxing, DLP/PII scanners for data hygiene), and a mapping to **concrete AWS services** where we already run workloads.

- At **Layer 2**, we secure how AI systems and agents are built and run using strong non-human identity, prompt and output guardrails, sandboxed runtimes, secured vector stores, and model supply-chain/drift controls—implemented primarily with AWS primitives such as IAM, API Gateway, WAF, Lambda, ECS Fargate/EKS, VPC/Security Groups, KMS, CloudTrail/CloudWatch, Macie/Comprehend, and ECR/CodeArtifact/Signer.
- At **Layer 3**, we run AI as a governed program with AI inventory and risk tiering, policy and standards, HITL checkpoints and kill switches, AI-specific incident response, transparency, and non-human identity governance—implemented with AWS Organizations, IAM + tags, Config, IAM Access Analyzer, Security Hub, Audit Manager, and an AI Security Posture Management (AI-SPM) platform such as Cranium for continuous discovery, control testing, and ISO 42001 / NIST AI RMF / EU AI Act evidence.

Detailed mappings from each control to method, generic tools, and AWS services are provided in **Appendix A (Layer 2 Technical Controls)** and **Appendix B (Layer 3 Governance Controls).**

WHY governance is critical: organizations are rapidly deploying AI but lag in visibility, governance, and identity discipline for AI agents and services. Without a minimal governance spine—inventory, risk decisions, standards, compliance alignment, monitoring, and NHI-aware IAM—Layers 1 and 2 degrade into control "theater" and AI risk cannot be meaningfully managed at the program level
83% of organizations lack comprehensive AI security visibility[3]. Gartner's 2026 cybersecurity trends identify "agentic AI oversight" and "AI agent IAM adaptation" as top priorities, with 77% of organizations already using AI in cybersecurity but governance lagging significantly[25].

Run AI as a governed program (with NHI IAM)

Increasingly, this governance layer will be **instrumented by AI Security Posture Management (AI-SPM)** platforms. These tools continuously discover AI assets (models, agents, prompts, pipelines), map permissions and data flows, detect shadow AI, and flag misconfigurations and over-privileged access in real time. When wired into identity and cloud controls, AI-SPM becomes the "runtime sensor" for Layer 3—proving that policies, NHI IAM standards, and cross-border rules are actually enforced in production.

Emerging challenge—AI agent identity management: AI agents represent a new class of non-human identities (NHI) requiring sophisticated lifecycle management. Traditional IAM frameworks struggle with:

- Dynamic access needs that change based on task context
- Lack of traceable human ownership for autonomous agents
- Audit and remediation challenges when agents act with elevated privileges
- Emergency access revocation when AI-specific risks emerge rapidly[38][42]

Organizations must extend identity governance to include standardized authentication for AI agents, least privilege enforcement, time-limited access (JIT provisioning), and comprehensive audit trails[38][42].

Alignment: NIST AI RMF (Map, Measure, Manage, Govern), NIST CSF 2.0 "Govern" function, SANS AI governance playbooks, Critical AI Security Guidelines[9][11][12][15][17].

Recommendations: A Sequenced Approach

Phase 1: Stabilize the Cyber Residual Risk Floor (Immediate)

1. **Audit and harden identity controls**
    - Enforce MFA across all user and admin accounts
    - Implement privileged access management (PAM) with session recording
    - Rotate credentials and eliminate standing privileged access
    - Deploy adaptive authentication based on risk signals
2. **Prioritize KEV catalog remediation**
    - Establish automated vulnerability scanning and tracking
    - Remediate CISA Known Exploited Vulnerabilities within SLA
    - Harden internet-facing assets and reduce attack surface
3. **Implement cloud security posture management**
    - Deploy CSPM tools for continuous misconfiguration detection
    - Enforce least privilege IAM policies and role-based access
    - Scan infrastructure-as-code for security issues pre-deployment
4. **Strengthen email and endpoint defenses**
    - Enable email authentication (SPF, DKIM, DMARC)
    - Deploy EDR with behavioral detection for AI-generated malware
    - Conduct AI-aware phishing simulations and training
5. **Establish logging and backup discipline**
    - Centralize logging with tamper-proof retention
    - Implement immutable backups with tested recovery procedures
    - Create and test incident response playbooks for AI-enabled attacks

Success metric: Achieve CIS Controls IG1 compliance baseline. Measure Mean Time to Detect (MTTD) and Mean Time to Respond (MTTR) for identity and vulnerability incidents.

AICM's role in Phase 1 is deliberately narrow: use it as a cross-check to confirm that your existing CSF/CIS/CCS hygiene covers the relevant "cloud-only" and "AI+cloud" control objectives, but resist the temptation to prioritize AI-specific controls before this floor is solid. For most organizations, the majority of AI-amplified loss is still removed by getting these foundational AICM-aligned controls to "good enough."

Phase 2: Secure AI Systems and Agentic Workflows (Near-Term)

1. **Implement OWASP Top 10 for LLM controls**

- Deploy input validation and output sanitization for all LLM interactions
- Isolate prompt context from system instructions
- Sandbox tool execution and enforce least privilege for tool access
- Monitor for prompt injection patterns and anomalous model behavior

2. **Establish AI supply chain security**
   - Verify provenance of models and training datasets
   - Use private model registries with artifact signing
   - Scan dependencies for known vulnerabilities
   - Sanitize public training data before use
3. **Design secure agentic architectures**
   - Require human-in-the-loop for high-impact decisions
   - Limit agent autonomy with explicit scope and time boundaries
   - Implement comprehensive audit logging for all agent actions
   - Integrate DLP to prevent sensitive data disclosure
4. **Deploy AI-specific monitoring and detection**
   - Monitor for model drift and unexpected outputs
   - Detect data exfiltration attempts via encoded URLs or side channels
   - Alert on excessive resource consumption (model DoS)
   - Track failed injection attempts and suspicious prompt patterns

Success metric: Achieve OWASP LLM security checklist compliance for production AI systems. Track prompt injection detection rate and AI-related security incidents.

In Phase 2, AICM becomes the main control catalog: prioritize the subset of AICM controls tagged as AI-specific in the model, data, and application/tool layers, and use OWASP Top 10 for LLMs as the threat lens to select what actually matters for your architectures. Instead of "implementing all 243 controls," pick the small number of AICM items that directly mitigate prompt injection, unsafe tools, poisoning, and agentic misuse in your highest-value workflows.

Phase 3: Operationalize AI Governance (Ongoing)

1. **Build AI inventory and risk classification**
   - Catalog all AI systems, models, and use cases
   - Classify by risk tier (high/medium/low) based on data sensitivity and decision impact
   - Assign clear ownership and accountability for each system
   - Track Shadow AI usage and bring unsanctioned tools under governance
2. **Establish AI risk management program**
   - Conduct AI-specific risk assessments aligned with NIST AI RMF
   - Document risk acceptance decisions with executive sign-off
   - Create pre-deployment security review process for new AI use cases
   - Maintain incident response plans for AI-specific breaches
3. **Define AI policies and standards**
   - Publish acceptable use policies for AI tools
   - Establish secure development lifecycle (SDL) requirements for AI
   - Set data handling standards (anonymization, retention, access controls)
   - Create ethical AI guidelines addressing bias, transparency, accountability
4. **Implement identity governance for AI agents**
   - Extend IAM framework to manage non-human identities (AI agents)
   - Enforce least privilege with time-limited, just-in-time (JIT) access
   - Deploy comprehensive audit trails for all AI agent activities
   - Establish emergency revocation procedures for compromised agents
5. **Map to regulatory requirements**
   - Align with sector-specific regulations (GDPR, HIPAA, financial services)

- Prepare for emerging AI-specific regulation (EU AI Act, state-level laws)
- Conduct periodic compliance audits with external validation
- Maintain documentation for regulatory inquiries and audits

6. **Continuous assurance and improvement**
    - Perform quarterly AI security assessments
    - Track AI security metrics (Shadow AI incidents, prompt injection attempts, agent privilege violations)
    - Conduct tabletop exercises for AI-enabled attack scenarios
    - Update policies and controls based on emerging threats and lessons learned

Success metric: Achieve comprehensive AI inventory coverage (>95% of systems cataloged). Reduce Shadow AI incidents by 80% within 12 months. Demonstrate regulatory compliance in external audits.

Phase 3 is where you can lean on AICM's governance, transparency, and compliance domains—plus the AI-CAIQ questionnaire built on top of AICM—to turn your inventory, risk reviews, and NHI IAM standards into something that can be audited, procured against, and shared with regulators or customers. The same AICM controls can be used internally for continuous assurance and externally as part of vendor due diligence and contractual expectations for AI services.

Conclusion: Risk-First AI Security

The AI security investment paradox—high spending yet low confidence—stems from misaligned priorities. Organizations investing in AI-specific controls while neglecting foundational cyber hygiene are building on sand. AI doesn't create new primary loss channels; it amplifies exploitation of identity weaknesses, misconfigurations, and vulnerabilities at industrial scale.

The path forward is sequential:

1. **Stabilize the cyber residual risk floor** using CSF 2.0 and CIS Controls to harden identity, patch management, cloud security, email/endpoint defenses, and logging/backup discipline (Layer 1)
2. **Secure AI-native architectures** by implementing OWASP Top 10 for LLMs, supply chain verification, agentic design patterns, and AI-specific monitoring (Layer 2)
3. **Govern AI as a program** with inventory, risk assessment, policy enforcement, regulatory alignment, identity governance for AI agents, and continuous assurance (Layer 3)

This three-layer model maps cleanly to NIST CSF 2.0, NIST AI RMF, OWASP guidance, and SANS AI security domains. It prioritizes investment based on actual loss drivers—$2.7-$6.3 billion in BEC losses[36][40], $4.88 million average phishing breach costs[36], 60% of attacked SMBs closing within six months[21][27]—rather than theoretical risks.

The message for security leaders is clear: Don't talk AI security until you've hardened the environment AI-assisted attackers will exploit, addressed AI-native failure modes, and wrapped it all in governance that ensures accountability, compliance, and continuous improvement.

AI security is cyber security with AI-specific extensions, not a separate discipline. Organizations that recognize this—and invest accordingly—will achieve the 6% "very capable" status that eludes most enterprises today.

4. Appendices

FIRST COMMON TOOLS / METHODS

Overall tools for Layer 2 and 3

Layer 2 – AI / App Security (technical)

- Methods: secure agent identity, prompts/I/O, tools, memory, and runtime.
- Tool types:
    - API & policy: API gateway, WAF, policy engine (RBAC/ABAC), schema validators.

        - Runtime: container/VM platform, network segmentation, logging/metrics, DLP/PII scanners, model registry.
        - Identity: IAM / WIAM for NHIs (agents, tools, pipelines).
    - AWS core: IAM, API Gateway/ALB, WAF, Lambda, ECS Fargate/EKS, VPC/Security Groups, KMS, CloudTrail, CloudWatch, Macie/Comprehend, ECR/CodeArtifact/Signer, DynamoDB/RDS/OpenSearch.

Layer 3 – Governance, NHI IAM, AI-SPM

- Methods: run AI as a program (inventory, risk, policy, oversight, IR, NHI governance).
- Tool types:
    - Governance: AI-SPM / AI governance platform (e.g., Cranium), GRC/IRM, IR/case management.
    - Identity governance: IGA/WIAM for non-human identities, access reviews, lifecycle, risk scoring.
    - Cloud posture: CSPM (plus Config/Security Hub) for continuous technical enforcement.
- AWS core: Organizations, IAM + tags, Config, IAM Access Analyzer, Security Hub, Audit Manager, CloudTrail Lake.
- AI-SPM: Cranium as the main overlay for AI inventory, AI-BoM, agent/tool graphs, control attestation, and ISO 42001/NIST/EU AI Act evidence.

Appendix A – Layer 2 Technical Controls

- Title: **Appendix A – Layer 2 Technical Controls (Method → Tools → AWS)**
- Content: the table that lists for each Layer-2 control:
    - Method (what "good" does)
    - Typical tools/patterns
    - AWS services (examples)

This includes the Top 5 + 2 and the rest of the technical controls (1–12, 16–18).

A.1 Controls 1–6, 7, 9, 10 (Top 5 + 2 covered)

Control	Method (what "good" does)	Typical tools / patterns	AWS services (examples)
1. Agent Identity & Auth	Treat every agent as a first-class non-human identity with unique credentials and strong auth; no shared keys.	IAM / IdP, OIDC/JWT or mTLS, secret management, tagging and lifecycle workflows.	IAM roles for agents (with NHI tags), IAM Identity Center or Cognito / custom OIDC for JWTs, STS for short-lived creds, AWS Private CA + mTLS, Secrets Manager.
2. Least-Privilege Tool Access	Expose tools as scoped APIs; bind each agent to minimal permissions per tool; regularly review.	API gateway, RBAC/ABAC policy engine, IAM policy analysis, access reviews.	API Gateway / ALB in front of tools, IAM role policies with least-privilege actions/resources, AWS Config managed/custom rules to block wildcards, IAM Access Analyzer to detect excessive or external access.
3. Agent-to-Agent Trust Boundaries	Require auth and message signing between agents; apply zero-trust even on internal buses.	mTLS / OIDC, signed messages, service mesh or API gateway.	Private CA + mTLS between services, Cognito/IdC-issued JWTs checked at API Gateway, App Mesh / service-to-service TLS, signed messages over SQS/SNS/Kinesis (validated by Lambda/orchestrator).

Control	Method (what "good" does)	Typical tools / patterns	AWS services (examples)
4. Prompt Injection Defense	Route all user/RAG/tool input through an input firewall that validates, classifies, and sanitizes before the model/tools.	Web app firewall, request validation, input sanitization, L7 classifiers/DLP, logging.	API Gateway request validation, AWS WAF (custom rules for injection patterns), Lambda "guardrail" layer, Amazon Comprehend for text classification, Macie for sensitive data patterns, CloudWatch Logs for telemetry.
5. Output Validation & Filtering	Enforce schemas and safety checks on model outputs before rendering or executing; never let raw output call tools directly.	JSON schema validation, policy engine, sandboxed execution, DLP checks.	Lambda / Step Functions to validate output schema and apply policy; API Gateway response validation; Comprehend/Macie for output content checks; ECS/Fargate/EKS sandboxes for any code/command execution derived from outputs.
6. PII Detection & Scrubbing	Detect and redact PII/secrets in inputs, memory, logs, and third-party calls; minimize data.	DLP/PII scanners, tokenization / masking, key management, logging.	Amazon Macie (S3 scans), Comprehend PII entity detection for prompts/RAG chunks, GuardDuty findings for suspicious exfil, Lambda filters to mask/tokenize, KMS for encryption, Secrets Manager for secret storage.
7. Action Boundaries & Guardrails	Define explicit allowlists per agent role (what actions on which systems/data) and enforce centrally.	Policy engine (RBAC/ABAC), capability catalogs, code review, unit tests.	Central "tool policy" in Lambda/OPA that all agent→tool calls pass through; IAM roles and resource policies encoding allowlists; Config rules to flag disallowed actions; API Gateway resource policies.
8. Rate Limiting & Throttling	Prevent runaway agents and cost spikes by enforcing quotas and throttling; alert on anomalies.	API rate limiting, quota management, monitoring, alerting, budgets.	API Gateway throttling and usage plans; ALB/WAF rate limits; CloudWatch metrics and alarms on call volume and latency; AWS Budgets / Cost Anomaly Detection for spend spikes.
9. Agent Sandboxing & Isolation	Run agents in isolated runtimes with minimal network/FS access and clear blast-radius boundaries.	Containers/VMs, network segmentation, no direct host access, hardened images.	ECS Fargate or EKS with private subnets, Security Groups restricting egress to approved endpoints, VPC endpoints for data stores, no public IPs for agent tasks, ECR for images, KMS-encrypted volumes and data.

Control	Method (what "good" does)	Typical tools / patterns	AWS services (examples)
10. Immutable Audit Trail Logging	Log every tool call and key decision with tamper-evident storage and rich context for investigation.	Central log pipeline, WORM storage, SIEM, correlation IDs.	CloudTrail for AWS API activity, CloudWatch Logs for app/tool logs, S3 log buckets with Object Lock + KMS, CloudTrail Lake / Athena for querying, Security Hub as aggregator of findings.
11. Agent Memory Security	Encrypt and segment vector/RAG stores; enforce access by tenant/use case; block direct queries.	Encrypted DB/vector store, fine-grained access control, network isolation.	OpenSearch, Aurora, DynamoDB, or other stores with KMS encryption and IAM-scoped access; VPC endpoints only; per-tenant or per-namespace IAM policies; no public S3/OpenSearch for memory.
12. Session & State Integrity	Bind session state to user+agent and validate each turn to detect poisoning or hijack.	Session store, integrity checks, replay detection, anomaly monitoring.	DynamoDB / ElastiCache for session state; Lambda/Step Functions layer to validate historical turns before reuse; CloudWatch metrics for unusual session patterns.
16. LLM Supply Chain Security	Control which models/datasets/plugins are allowed; verify provenance and pin versions.	Model registry, artifact signing, SBOM/AI-BoM, dependency scanning.	ECR / CodeArtifact for model/tool images, AWS Signer for artifact signing, Parameter Store or DynamoDB for "approved model list" and versions, Config rules to block unapproved repos/regions.
17. Hallucination & Reliability Controls	Ground high-impact outputs in systems of record; enforce confidence thresholds and fallbacks.	RAG patterns, validation against SoRs, confidence scoring, HITL for low confidence.	Orchestrator logic in Lambda/ECS that always validates decisions against RDS/DynamoDB/other SoRs; CloudWatch metrics for error/refusal/confidence distributions; Step Functions/HITL when below thresholds.
18. Model Behavioral Drift Monitoring	Track output distributions, error/refusal rates, and key metrics over time; alert on shifts.	Telemetry pipeline, statistical monitoring, anomaly detection.	CloudWatch custom metrics for model behavior (e.g., classification ratios, refusals), CloudWatch Alarms, Kinesis Firehose / Lambda for streaming logs to analytics, QuickSight or external lakehouse for trend analysis.

structure: overall **methods** and **tool patterns**, with **AWS services** as the realization for each AI control

Appendix B – Layer 3 Governance Controls

- Title: **Appendix B – Layer 3 Governance Controls (Method → Tools → AWS + AI-SPM)**
- Content:
    - Table for 13–15 (HITL, kill switch, rollback).

- Table for 19–20 (IR and transparency).
- Table for NHI IAM & AI-SPM (identity governance and Cranium overlay).

These tables show, for each governance control, the method, generic tools (IR platform, IGA, GRC, AI-SPM), and AWS/Cranium mappings.

Appendix B – Layer 3 Governance Controls

B.1 Oversight and safety (13–15)

Control	Method (what "good" does)	Typical tools / patterns	AWS services (examples)
13. Human-in-the-Loop Checkpoints	Gate irreversible or high-impact actions through human approvals with full audit trails.	Workflow engine, approval UI, ticketing/ITSM, logging.	Step Functions to orchestrate flows and pause for approval; Lambda + DynamoDB for approval state; internal UI via AppRunner/API Gateway; CloudTrail/CloudWatch Logs for who/what/when.
14. Kill Switch & Emergency Halt	Provide fast per-agent, per-session, and global halt mechanisms, tested regularly.	Feature flags, routing controls, monitoring/alerts, playbooks.	AppConfig or DynamoDB flags checked by all agents; Route 53 / API Gateway stage/route switching for global cutover; CloudWatch Alarms + SNS for triggering and alerting; runbooks in your IR tooling.
15. Rollback & Action Reversibility	Design actions with compensating steps and logs sufficient to undo changes.	Transaction logs, compensating workflows, backup/restore, change management.	RDS/DynamoDB PITR; AWS Backup policies; CloudTrail Lake for reconstructing sequences; Step Functions for compensating workflows (e.g., reverse payment, revert config).

B.2 Compliance, IR, transparency (19–20)

Control	Method (what "good" does)	Typical tools / patterns	AWS services (examples)
19. AI Agent Incident Response Plan	Extend IR to AI-specific incidents (harmful output, unauthorized action, runaway agent, shadow AI), with clear notification paths.	IR platform, SIEM, case management, runbooks, tabletop exercises.	Security Hub as central finding store; GuardDuty, Config, CloudTrail as signal sources; CloudWatch Alarms + SNS for paging; Audit Manager for documenting control coverage and response evidence.
20. Agent Transparency & Explainability	Make AI use visible and record "why" for high-risk decisions in human-readable form.	UX labels, explanation generation, decision logging, report templates.	Application layer on Lambda/ECS to generate explanations stored in DynamoDB/CloudWatch; correlation IDs in CloudTrail/CloudWatch; static assets in S3 for transparency notices on front-ends.

B.3 NHI IAM & AI-SPM (cross-cutting for Layer 3)

Area	Method (what "good" does)	Typical tools / patterns	AWS + AI-SPM services (examples)
NHI IAM standard (agents as identities)	Treat agents and other NHIs as first-class identities with ownership, purpose, risk tier, lifecycle, and least-privilege access.	IGA / non-human identity governance, WIAM, tagging standards, certification workflows.	IAM roles (per agent/tool) with mandatory NHI tags (Owner, Purpose, RiskTier, DataClass, Expiry); AWS Organizations tag policies; Config rules to enforce tags and deny wildcards; IAM Access Analyzer for over-privilege and external exposure.
NHI lifecycle & reviews	Ensure NHIs are created via governed workflows, reviewed, and retired; focus reviews on high-risk NHIs.	Identity lifecycle engine, access review/certification, risk-based prioritization.	Config rules + Access Analyzer + CloudTrail Lake to identify stale/over-privileged roles; Security Hub insights for "agent role without owner/overdue rotation"; external IGA/WIAM for review UI and campaigns.
AI-SPM / governance plane	Continuously discover AI assets (models, agents, pipelines), map dependencies, test controls, and generate compliance artifacts.	AI-SPM / AI governance platform, CSPM/DSPM, GRC.	Cranium AI Security Platform for: auto-discovery of AI systems, AI-BoMs and AI Cards, control attestation against ISO 42001 / NIST AI RMF / EU AI Act, drift/supply-chain monitoring, and reporting; AWS Config/Security Hub/CloudTrail as data sources feeding Cranium.

Overall Layer-2/3 tooling suggestions (beyond the Top-20)
Top-20 mostly covers the *what*; Layer 2 and 3 also need some program-level tools beyond individual controls:

- For Layer 2 (technical surface):
  - An **agentic orchestration framework** you control (where you can embed all these checks).
  - A **central policy/guardrail service** (could be a custom Lambda/OPA) that sits in front of all tools and agents.
  - Your existing **CSPM + container security** stack to keep the runtime sane (CIS, EKS/Fargate hardening, image scanning).
- For Layer 3 (governance):
  - An **AI-SPM / AI governance tool** (Cranium or similar) for inventory, AI-BoM, risk scoring, testing, and compliance mapping.
  - A **non-human identity governance** capability (could be added to your IGA) to enforce the NHI IAM standard across AWS, CI/CD, SaaS.
  - A **GRC / IR platform** to connect AI risks and incidents into your broader NIST CSF / ISO 42001 / NIST AI RMF program.

NOTE - the main doc keeps the three-layer story and the Top-20 mapping, while Appendices A and B carry the "how" with methods, generic tools, and AWS specifics.

Appendix C: Top-20 AI Agent Controls → Layer → AICM domain

This appendix links each of the Top-20 agent controls to the primary layer in our three-layer model and to the most relevant AICM domain(s), so architects and auditors can move directly from “what really matters” to specific AICM rows and evidence expectations

Mapping Top-20 Agent Controls to AICM

#	Top-20 AI agent control	layer	Likely AICM domain(s) (CSA AICM)
1 *	Agent Identity & Auth	2	Identity & Access Management; Non-Human Identity Management; Technical Controls for Access
2	Least Privilege Tool Access	2	Identity & Access Management; Application & Tool Security; Data Access Governance
3	Agent-to-Agent Trust Boundaries	2	Identity & Access Management; Network & Infrastructure Security; Secure Service Communication
4 *	Prompt Injection Defense	2	Application & Tool Security; Model Security & Robustness; Data Security & Privacy (input controls)
5	Output Validation & Filtering	2	Application & Tool Security; Safety & Content Governance; Data Security & Privacy (output controls)
6 *	PII Detection & Scrubbing	2	Data Security & Privacy; Records Management; Regulatory & Privacy Compliance
7	Action Boundaries & Guardrails	2	Application & Tool Security; Safety & Reliability; Operational Controls for Autonomy
8	Rate Limiting & Throttling	2	Infrastructure & Operations Security; Availability & Resilience; Abuse & DoS Protection
9 *	Agent Sandboxing & Isolation	2	Infrastructure & Environment Security; Network Segmentation & Isolation; Workload Protection
10 *	Immutable Audit Trail Logging	2 (tech) / 3 (evidence)	Logging & Monitoring; Audit & Accountability; Forensics & Investigations
11	Agent Memory Security	2	Data Security & Privacy; Storage & Encryption; Access Control for AI Context Stores
12	Session & State Integrity	2	Application & Session Management; Integrity & Anti-Tampering Controls
13 *	Human-in-the-Loop Checkpoints	3	Human Oversight & Control; Safety & Risk Management; Governance & Accountability
14 *	Kill Switch & Emergency Halt	3	Incident Management & Response; Safety Controls; Operational Resilience

#	Top-20 AI agent control	layer	Likely AICM domain(s) (CSA AICM)
15	Rollback & Action Reversibility	3	Change & Configuration Management; Recovery & Continuity; Incident Response & Remediation
16	LLM Supply Chain Security	2 (tech) / 3 (governance)	AI Supply Chain Management; Model & Dataset Provenance; Third-Party & Vendor Risk
17	Hallucination & Reliability Controls	Layer 2	Safety & Reliability; Model Evaluation & Monitoring; Application & Tool Security
18	Model Behavioral Drift Monitoring	2 (tech) / 3 (assurance)	Model Monitoring & Drift; Performance & Risk Metrics; Continuous Assurance
19	AI Agent Incident Response Plan	3	Incident Management & Response; Governance & Risk Management; Regulatory & Communications Management
20	Agent Transparency & Explainability	3	Transparency & Explainability; User Communication & Disclosure; Ethics & Accountability

seven must-have controls with justifications:

1 - #1 Agent Identity & Auth – Makes every agent a governed non-human identity, so you can actually enforce least privilege, track actions, and revoke access when things go wrong.
2 - #4 Prompt Injection Defense – Cuts off the primary exploited vector for agents by stopping untrusted prompts and retrieved content from directly steering tools and data access.
3 - #6 PII Detection & Scrubbing – Prevents accidental privacy and regulatory landmines by stopping sensitive data from leaking into prompts, memory, logs, and third-party APIs.
4- #9 Agent Sandboxing & Isolation – Limits blast radius by ensuring agents run in tightly scoped environments that can't freely reach your network, file systems, or data stores.
5- #10 Immutable Audit Trail Logging – Gives you the forensic spine—who did what, when, with which tools—so you can investigate incidents, prove compliance, and continuously improve.
6 - #13 Human-in-the-Loop Checkpoints – Inserts human judgment before irreversible or high-impact actions, turning many catastrophic failures into simple near-misses.
7- #14 Kill Switch & Emergency Halt – Guarantees you can stop misbehaving agents fast at agent, session, or global scope, which is essential for containing emerging AI incidents.

Appendix C– Layer 3 Governance Deep Dive (Technical, Operational, Cross-Border)

This appendix collects the richer breakdown and rationale so the main body stays balanced.

A. Expanded Layer 3 structure

Layer 3 can be decomposed into three complementary dimensions:

- **Technical governance** – architectures, standards, NHI IAM, and technical assurance.
- **Operational governance** – ownership, processes, risk decisions, and lifecycle execution.
- **Cross-border governance** – regulatory, jurisdictional, and data-movement aspects, including third-party ecosystems.

This structure lets you keep the main body concise while giving practitioners enough detail to design concrete programs.

B. Technical governance – architectures, standards, and NHI IAM

Architectures and control patterns

- Define reference patterns for RAG, agentic orchestration, and tool integration that bake in mitigations for AI/LLM failure modes (prompt injection, insecure tools, data leakage, model DoS, supply chain issues).
- Explicitly map each pattern to NIST CSF 2.0 and AI-specific guidance so AI is treated as an extension of existing security architecture, not a parallel stack.

Data and model governance

- Set rules for training data and RAG corpus selection, retention, and provenance (what data is allowed, under what basis, and where it may reside).
- Implement output-side controls (filters, policy-aware response guards) and logging/alerting to detect and contain sensitive data exfiltration through prompts and tool outputs.

NHI IAM as a core standard

- Introduce a **standardized NHI model** (service principals, workloads, AI agents, connectors, schedulers) with mandatory attributes: owner, purpose, environment, data-access class, risk tier, and expiry/review dates.
- Enforce **lifecycle discipline**: NHIs created only via approved workflows, always tied to a human owner and system of record; automatic detection and cleanup of orphaned identities and secrets.
- Apply **least privilege and segmentation**: pre-defined role patterns for AI agents, scoped to minimal resources and environments; default-deny for cross-tenant/cloud/region access with explicit approvals.
- Run **continuous assurance**: automated key/secret rotation, behavioral anomaly detection for NHIs, and periodic access recertification with policy that disables non-attested identities.

We proposed a concrete NHI IAM standard to NIST NCCoE for consideration
https://docs.google.com/document/d/11BV7NbwhF9drZZiPm9hlw7RJebMbdhpH/edit

Technical assurance and supply chain

- Embed AI/agentic concerns in the SDLC: threat modeling, abuse-case testing, evaluation harnesses, and (where appropriate) red-team exercises focused on AI misuse and tooling abuse.
- Treat models, datasets, and plugins as supply-chain artifacts: require provenance, integrity verification, and monitoring for poisoned or compromised components.

AICM can serve as the technical checklist behind this section: its Model Security, Data Security & Privacy, AI Supply Chain, and Infrastructure controls provide concrete requirements for RAG and agentic reference patterns, NHI IAM standards, and supply-chain assurance. Rather than drafting your own AI-specific control library, you can select and profile AICM control IDs to match your architectures and treat them as the authoritative "what good looks like" for technical AI governance

C. Operational governance – ownership, decisions, lifecycle

Program ownership and decision rights

- Name an AI risk/program owner and establish a simple RACI across security, data, legal, and business functions.
- Define risk categories and decision thresholds so local teams know which use cases they can approve and which must be escalated.

Use-case intake and approval

- Standardize intake templates capturing purpose, data, models/providers, NHI requirements, risk category, and expected impact.
- Use tiered review: fast-path for low-risk internal productivity use; structured review for customer-facing, safety-relevant, or compliance-sensitive AI.

Operations and incident handling

- Develop playbooks for AI-specific incidents (prompt injection, agent overreach, AI-assisted fraud, model degradation) with clear containment and rollback paths.
- Run training for engineers and business users on safe AI usage, including how to avoid and report Shadow AI.

Lifecycle and metrics

- Periodically re-evaluate AI systems as models, data, and regulations evolve; have criteria for retraining, reconfiguration, or retirement.
- Track key metrics (AI incidents, exception requests, NHI IAM hygiene, Shadow AI findings) and report them into leadership and board-level risk views.

On the operational side, AICM's lifecycle tags (preparation, development, evaluation, deployment, operations, retirement) can be overlaid on your intake, review, and reassessment processes so every AI use case knows which control objectives must be demonstrated at each stage. This turns abstract governance principles into concrete, time-bound tasks for product teams, security, and risk owners.

D. Cross-border governance – regulation, data movement, ecosystem

Regulatory and policy alignment

- Map AI use cases to applicable regulations and internal policies and label them by regulatory sensitivity.
- Encode region-specific constraints into technical enforcement: data residency rules, allowed model locations, and routing logic for region-bound data.

Third-party and ecosystem risk

- Define due-diligence requirements for external AI services and APIs, including their own data handling, logging, NHI/IAM posture, and incident commitments.
- Use contracts and technical controls (e.g., minimization, masking, tenant isolation) to constrain cross-border and partner data flows.

Cross-border NHI governance

- Set and enforce where NHIs may run and what regional data they can access (identity policy, IAM conditions, segmentation).
- For NHIs that operate across clouds, tenants, or partners, clearly document shared responsibility and trust boundaries and reflect that in technical design (separate identities per boundary, scoped tokens, independent logs).

For cross-border and third-party risk, AICM's data, privacy, and compliance controls can be used as a standard clause set: you can reference specific AICM control IDs in contracts, DPAs, and vendor questionnaires to require minimum AI data-handling, residency, logging, and incident-handling practices from providers. This keeps your cross-border governance consistent with the same control framework you apply internally, instead of maintaining a separate "vendor-only" checklist.

Appendix X – Quantitative Perspective on "What Really Matters" in AI Security

This appendix provides a directional, data-informed assessment of how much each layer of the AI security model contributes to **realistic AI risk reduction in expected-loss terms**. The goal is not to produce precise actuarial numbers, but to anchor priorities in where current evidence shows AI actually impacts organizations.

X.1 Framing: AI as an amplifier of existing loss paths

Across recent threat-intelligence reports and breach-cost studies, three patterns are consistent:

- AI is heavily used to **amplify traditional vectors**—phishing/BEC, credential stuffing, reconnaissance, vulnerability exploitation, and ransomware—rather than introducing wholly new primary loss channels at scale.
- The largest AI-related cost premiums come from **shadow AI and governance gaps** layered on top of those same technical weaknesses (for example, unmanaged AI tools handling sensitive data without proper access controls).
- AI-native failures—prompt injection, unsafe tools, AI-generated code flaws, over-delegated agents—are real and growing, but still represent a minority of all incidents compared to classic phishing/credential/ransomware paths.

Given that empirical backdrop, we can ask: *If an organization were to push each layer of the model toward "good enough," what fraction of AI-driven loss would that realistically mitigate?*

X.2 Layer 1 – Cyber Residual Risk Floor (~60–70%)

Scope: Identity/MFA, privileged-access discipline, KEV-driven patching, exposure and misconfiguration management, backups/recovery, logging, and basic email/endpoint defenses.

Why it dominates:

- AI-enabled phishing and social-engineering volume is exploding, but successful incidents still rely on the same weaknesses: users without MFA, over-privileged accounts, exposed services, unpatched vulnerabilities, and flat networks. Hardening this "floor" removes the oxygen AI needs to convert its advantages into breaches.
- Ransomware and data-theft campaigns increasingly incorporate AI for target selection, phishing content, and automation, yet their ability to encrypt or exfiltrate at scale is gated by whether core controls (identity, patching, segmentation, backups) are in place.

Directional estimate:

For a typical enterprise IT or cloud-centric SMB environment, a strong Layer 1 posture likely delivers **on the order of 60–70% of total achievable AI-risk reduction** (in expected-loss terms). Most of the reduction comes from:

- Preventing AI-crafted phishing from yielding usable credentials.
- Denying AI-driven recon access to exploitable misconfigs and unpatched services.
- Limiting blast radius and recovery time when AI-accelerated ransomware or intrusion does occur.

This is why the "foundations first" narrative is not just philosophically attractive—it is **mathematically dominant** for AI, just as it is for non-AI attacks.

X.3 Layer 2 – AI / Application Security (~15–25%)

Scope: AI-specific and agentic patterns: prompt injection, insecure output handling, unsafe tool integration, model and data poisoning, AI-generated code risks, and excessive or poorly-governed agent autonomy.

Where it moves the needle:

- **Confidentiality:** Prompt-injection and "reprompt"-style attacks can cause models to disclose sensitive internal data, query backends in unintended ways, or misuse tools; securing context assembly, outputs, and tools significantly cuts this risk even in well-hardened environments.
- **Integrity:** AI-generated code and autonomous agents can introduce vulnerabilities or mis-configure systems at scale if not constrained by review and guardrails.
- **Availability and abuse:** Model DoS, resource abuse, and agents looping over critical systems can degrade services that have become operationally important.

These issues are increasingly visible in AI-incident catalogs and case studies, but they still represent a **smaller share of volume and dollar loss** than the fundamental identity/phishing/ransomware axis in most organizations.

Directional estimate:

Once Layer 1 is competent, a mature Layer 2 capability plausibly adds **another ~15–25% of AI-risk reduction**, primarily by:

- Preventing AI-mediated data exfiltration and misuse of internal tools.
- Avoiding self-inflicted incidents where AI code or agents create new exploitable surfaces.

Layer 2 is where AI stops being "just" an attacker amplifier and becomes a defended, first-class application surface.

X.4 Layer 3 – Governance, Shadow AI, and NHI IAM (~10–20%)

Scope: AI inventory and classification, AI-specific risk assessment and acceptance, shadow-AI discovery and control, non-human identity (NHI) governance for agents and services, cross-border/data-residency controls, and third-party AI risk management.

Risk profile it addresses:

- **Shadow AI & oversight gaps:** Studies show AI-related breaches are significantly more expensive when unsanctioned tools and weak access controls are involved. Governance does not change

whether a phishing email arrives, but it strongly influences whether sensitive data ends up in unmanaged AI systems and how hard it is to detect and respond when it does.

- **Regulatory and cross-border exposure:** As AI touches regulated data and crosses jurisdictions, the presence or absence of programmatic governance often determines whether an incident becomes a routine breach or a high-profile regulatory event.
- **NHI sprawl:** Poorly governed non-human identities—agents, services, integrations—can quietly re-open powerful access paths even after Layer 1 and 2 controls are in place.

These risks are **lower frequency by count**, but often **higher severity** when they occur, especially for regulated or cross-border environments.

Directional estimate:
For a "typical" organization (not a hyperscaler or highly regulated giant), a credible Layer 3 program is likely responsible for **~10–20% of additional AI-risk reduction**, with an emphasis on:

- Reducing the cost and downstream impact of AI-related breaches (rather than completely preventing them).
- Flattening the tail of catastrophic, board-level AI incidents (regulatory fines, systemic misuse, reputational damage).

In other words, Layer 3 does not compete with Layer 1 on volume, but it is **disproportionately important for survivability and trust**.

X.4.1 AI Security Posture Management (AI-SPM) as Layer-3 tooling
A new class of **AI Security Posture Management (AI-SPM)** tools has emerged to operationalize Layer-3 governance. Common capabilities include:

- Continuous discovery of AI models, agents, integrations, notebooks, and AI SaaS usage (shadow AI detection).
- Posture assessment for AI assets: misconfigurations, internet exposure, risky data access paths, unsafe prompts/toolchains.
- Identity-centric analysis for AI agents and other NHIs—over-privilege detection, stale credentials, and automated right-sizing of permissions.
- Policy and compliance enforcement mapped to NIST AI RMF / CSF "Govern," including evidence for audits and AI-specific control frameworks.

In practice, AI-SPM is the **control plane and telemetry layer** for governance: it closes the loop between design-time policies (inventory, NHI IAM standards, cross-border rules) and runtime reality across cloud, SaaS, and agentic environments.

X.5 Putting the three layers together
A concise way to summarize the relative contribution of each layer is:

- **Layer 1 (Residual cyber risk floor):** Delivers the majority of AI-risk reduction (roughly two-thirds), because AI largely rides the same identity, phishing, vuln, and ransomware rails as non-AI attacks.
- **Layer 2 (AI / app security):** Adds a meaningful second tranche of risk reduction (roughly one-fifth), focused on AI-specific surfaces—how models, tools, and agents can be abused even in hardened environments.
- **Layer 3 (Governance, shadow AI, NHI IAM, cross-border):** Contributes the final tranche (roughly one-tenth to one-fifth), less visible in incident counts but crucial in shaping the **severity profile** and long-term acceptability of AI.

This framing supports a simple prioritization message for the main body of the paper:

1. **Do not treat AI as a separate planet.** First, reduce the cyber residual risk floor until AI-boosted attacks have very little to grab onto.
2. **Then secure AI as an application and agent layer.** Make sure AI itself is not the new biggest vulnerability.

3. **Finally, run AI as a governed program.** Use inventory, risk decisions, NHI IAM, and cross-border controls to keep the remaining AI risk within your appetite and regulatory boundaries over time.

***If you adopt AICM**, one practical approach is to tag your selected Layer 1, 2, and 3 controls with their estimated contribution to AI-risk reduction in your environment and focus first on the small subset of AICM controls that sit on the steepest part of that curve. This keeps the framework from turning into a 243-line compliance spreadsheet and preserves the risk-first story at the heart of this mode*

Appendix Y – AICM Profile for This Three-Layer Model

This appendix shows how to use the AI Controls Matrix (AICM) pragmatically: start with a small, high-value subset of domains and control types mapped to the three layers in this paper, instead of trying to "do all 243 controls."

Y.1 Layer-to-AICM profile table

Purpose: Give teams a starter "shopping list" of AICM focus areas per layer. You can localize to your own control IDs and profiles.

Model layer	AICM primary domains to emphasize	AICM control types / tags to prioritize	Typical usage in this model
Layer 1 – Cyber residual risk floor	Identity & Access, Infrastructure, Logging & Monitoring, Business Continuity & Resilience, Data Security (foundational)	Cloud-only and AI+cloud controls; core CCM-heritage controls	Use as a cross-check on your existing CSF / CIS / CCS posture; only add AI-specific work after this floor is stable.
Layer 2 – Fundamental AI / app security	Model Security, AI Supply Chain, Application & Tools Security, Data Security & Privacy Lifecycle	AI-specific controls; controls tagged to application/data layers	Select controls that directly mitigate prompt injection, unsafe tools, poisoning, and agentic misuse in your highest-value workflows.
Layer 3 – Governance, shadow AI, NHI IAM	Governance, Risk & Compliance (GRC), Transparency & Accountability, Data & Privacy, Third-Party Risk	Governance controls with lifecycle tags and SSRM ownership tags	Use as the backbone for AI inventory, risk acceptance, NHI IAM standards, cross-border rules, and vendor due diligence.

Y.2 Suggested "starter set" per layer

You will fill in exact AICM control IDs (e.g., AICM-MS-01) once you finalize your profile; below is the pattern of what to pick.

Layer 1 – Residual cyber risk floor

Pick a small set (for example, 8–12) of AICM controls that align with:

- Strong MFA, privileged access management, and least-privilege IAM for humans and basic NHIs.
- KEV-driven patching and misconfiguration management for cloud workloads and exposed services.
- Central logging, backup integrity, and tested recovery.
- Email and endpoint protections that reduce the success rate of AI-crafted phishing and ransomware.

Layer 2 – AI / application security
Pick a focused set (for example, 10–15) of AI-specific controls aligned with OWASP LLM Top 10 and your top loss drivers:

- Prompt and context controls: input validation, output filtering, context isolation, abuse-case testing.
- Tool and agent safety: least-privilege tool integrations, sandboxed execution, audit logging of agent actions.
- Model and data supply chain: provenance, artifact signing, private registries, poisoning detection or review.
- AI-specific monitoring: detection of prompt injection, model drift, anomalous outputs, and resource abuse.

Layer 3 – Governance, shadow AI, NHI IAM
Select a minimal but complete governance spine (for example, 8–12 controls) focused on:

- AI inventory and classification with risk tiers, ownership, and Shadow AI discovery.
- AI-specific risk assessment, risk acceptance, and pre-deployment review.
- Non-human identity governance for AI agents and services (lifecycle, least privilege, audit, emergency revocation).
- Cross-border and third-party AI risk management, including contractual expectations and AI-specific incident commitments.

Y.3 Ownership and lifecycle overlay
For every chosen AICM control, this model recommends tagging two attributes explicitly:

- **Owner:** CSP, model provider, orchestrated service provider, internal platform/app team, or security/governance function, using AICM's Shared Security Responsibility Model as the baseline.
- **Lifecycle stage:** preparation, development, evaluation, deployment, operations, or retirement, using AICM lifecycle tags so teams know *when* each control actually needs to be demonstrated.

This profile keeps AICM firmly in a supporting role: it is the control grid under a risk-first, three-layer investment story, not a 243-line checklist that drives priorities on its own.

References

[1] UK National Cyber Security Centre. (2024, January 23). The near-term impact of AI on the cyber threat. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat
[3] Kiteworks. (2025, August 20). 2025 AI Security Gap: 83% of Organizations Flying Blind. https://www.kiteworks.com/cybersecurity-risk-management/ai-security-gap-2025-organizations-flying-blind/
[4] Center for Security and Emerging Technology. (2024, November 18). Cybersecurity Risks of AI-Generated Code. https://cset.georgetown.edu/publication/cybersecurity-risks-of-ai-generated-code/
[7] Securiti. (2024, June 26). OWASP Top 10 for LLM Applications - Complete Guide. https://securiti.ai/owasp-top-10-for-llms/
[8] TechBuzz.ai. (2026, March 2). Google Exposes AI Weaponization in Cyber Attack Wave. https://www.techbuzz.ai/articles/google-exposes-ai-weaponization-in-cyber-attack-wave
[9] SANS Institute. (2026, March 3). Own AI Securely with SANS. https://www.sans.org/mlp/ai-security-blueprint
[10] Industrial Cyber. (2024, September 19). ENISA Threat Landscape 2024 identifies availability, ransomware, data attacks as key cybersecurity concerns. https://industrialcyber.co/reports/enisa-threat-landscape-2024-identifies-availability-ransomware-data-attacks-as-key-cybersecur
[11] NIST. (2024, September 18). Managing Cybersecurity and Privacy Risks in the Age of Artificial Intelligence.

https://www.nist.gov/blogs/cybersecurity-insights/managing-cybersecurity-and-privacy-risks-age-artificial-intelligence
[12] NIST. (2024, February 25). The NIST Cybersecurity Framework (CSF) 2.0. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf
[13] Oligo Security. (2025). OWASP Top 10 LLM, Updated 2025: Examples and Mitigation Strategies. https://www.oligo.security/academy/owasp-top-10-llm-updated-2025-examples-and-mitigation-strategies
[14] Evrim Agaci. (2025, October 16). Microsoft Warns AI Escalates Global Cyberattacks. https://evrimagaci.org/gpt/microsoft-warns-ai-escalates-global-cyberattacks-510668
[15] AI Governance Library. (2025, September 21). Critical AI Security Guidelines, v1.2. https://www.aigl.blog/critical-ai-security-guidelines-v1-2/
[17] Cranium.ai. (2026, January 11). AI Security in 2026: Enterprise Governance, Risks, and Best Practices. https://cranium.ai/resources/blog/ai-safety-and-security-in-2026-the-urgent-need-for-enterprise-cybersecurity-governance/
[21] Network Doctor. (2025, June 23). Massive AI Cyberattacks Cost SMBs $10.5T in 2025. https://www.networkdr.com/cybersecurity/ai-cyberattacks-smb-losses-2025-protect/
[22] Infosecurity Magazine. (2025, September 30). AI Tops Cybersecurity Investment Priorities, PwC Finds. https://www.infosecurity-magazine.com/news/ai-cybersecurity-investment-pwc/
[23] IBM. (2024, October 24). What Is Shadow AI? https://www.ibm.com/think/topics/shadow-ai
[24] Datadog. (2025, August 17). Abusing supply chains: How poisoned models, data, and third-party artifacts threaten AI applications. https://www.datadoghq.com/blog/detect-abuse-ai-supply-chains/
[25] LinkedIn. (2026, February 5). Gartner and Forrester Predict Quantum Security Spending to Rise in 2026. https://www.linkedin.com/posts/javier-galindo-ceo_gartner-identifies-the-top-cybersecurity-activity-7425569725608333312-hMGD
[26] Vanta. (2025, December 15). Top 6 AI security trends for 2026—and how companies can prepare. https://www.vanta.com/resources/top-ai-security-trends-for-2026
[27] The Network Installers. (2025, December 1). AI Cyber Threat Statistics: The 2025 Landscape of AI-Powered Attacks. https://thenetworkinstallers.com/blog/ai-cyber-threat-statistics/
[28] PwC. (2025, September 30). AI emerges as the top cybersecurity investment priority for companies in a shifting risk landscape. https://www.pwc.com/gx/en/news-room/press-releases/2025/pwc-digital-trust-insights.html
[29] ISACA. (2025, September 25). The Rise of Shadow AI: Auditing Unauthorized AI Tools in the Enterprise. https://www.isaca.org/resources/news-and-trends/industry-news/2025/the-rise-of-shadow-ai-auditing-unauthorized-ai-tools-in-the-e
[30] Australian Cyber Security Centre. (2025, October 15). Artificial intelligence and machine learning: Supply chain risks and mitigations. https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/artificial-intelligence-and-machine-learning-
[36] DeepStrike. (2025, April 28). Phishing Statistics 2025: AI, Behavior & $4.88M Breach Costs. https://deepstrike.io/blog/Phishing-Statistics-2025
[37] PurpleSec. (2025, December 6). Data Exfiltration Via AI Prompt Injection. https://purplesec.us/learn/data-exfiltration-ai-prompt-injection/
[38] Okta. (2025, July 13). Why Identity Governance for AI Agents is Your Next Big Challenge. https://www.okta.com/blog/industry-insights/beyond-human-users-why-identity-governance-for-ai-agents-is-your-next-big-challenge/
[39] IBM. (2025, January 16). How to calculate your AI-powered cybersecurity's ROI. https://www.ibm.com/think/insights/how-to-calculate-your-ai-powered-cybersecurity-roi
[40] Logstail. (2025, November 20). Phishing Report 2025: How AI, BEC and Human Error Fuel Billions in Cybercrime. https://logstail.com/blog/phishing-report-2025-how-ai-bec-and-human-error-fuel-billions-in-cybercrime/

[41] The Hacker News. (2026, January 14). Researchers Reveal Reprompt Attack Allowing Single-Click Data Exfiltration from Microsoft Copilot.
https://thehackernews.com/2026/01/researchers-reveal-reprompt-attack.html
[42] IDS Alliance. (2025, April 28). Identity and Access Management in the AI Era: 2025 Guide.
https://www.idsalliance.org/blog/identity-and-access-management-in-the-ai-era-2025-guide/
[44] Point Predictive. (2025, January 5). Emerging Threats of AI-Enabled Fraud in 2025.
https://pointpredictive.com/emerging-threats-of-ai-enabled-fraud/
[45] HackerOne. (2024, April 28). How a Prompt Injection Vulnerability Led to Data Exfiltration.
https://www.hackerone.com/blog/how-prompt-injection-vulnerability-led-data-exfiltration