The AI Security Ownership Crisis

How Structural Governance Failures Amplify Enterprise AI Risk

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-27

Table of Contents

Executive Summary

Enterprise AI adoption has entered a phase of structural instability. Organizations are embedding artificial intelligence into critical business operations at a pace that outstrips the governance, accountability, and security infrastructure required to manage that deployment safely. The result is not a technical vulnerability in the conventional sense — it is an institutional failure, a systematic absence of clear ownership over AI security that creates exploitable conditions at every layer of the stack, from ungoverned employee-facing tools to autonomous agents executing multi-step workflows against core enterprise systems.

The evidence is not anecdotal. Seventy-three percent of organizations report internal conflict over who owns AI security controls [1]. 26 percent have comprehensive AI security governance policies in place [2]. Seventy-six percent identify shadow AI — unauthorized deployment of AI tools by employees or business units without IT or security oversight — as a definite or probable problem, up 15 percentage points from the prior year [1]. Ninety-six percent of CISOs have been assigned responsibility for AI governance on top of their existing mandates [3], yet CISOs rank fourth in actual AI security decision-making authority within their organizations, behind CIOs, CTOs, and business unit leaders [1]. Just five percent of security leaders feel prepared to detect and contain a compromised AI agent [4]. Forrester analysts predict that 2026 will see the first publicly disclosed data breach attributable to an agentic AI deployment [5].

These findings converge on a single diagnosis: the AI security ownership crisis is not a gap to be closed by adding one more control or assigning one more policy owner. It is a structural failure requiring coordinated remediation across governance design, accountability frameworks, identity infrastructure, and organizational culture. This paper documents the anatomy of that failure, examines the specific structural patterns that sustain it, and provides a framework for remediation grounded in CSA's AI Controls Matrix (AICM), MAESTRO threat modeling, Zero Trust architecture, and the emerging body of agentic AI security guidance from the CSA AI Safety Initiative.

---

1. Introduction and Background

1.1 The Velocity Problem

AI adoption in the enterprise accelerated dramatically through 2025. By the end of that year, 88 percent of organizations reported using AI in at least one business function, up from 78 percent the prior year, yet approximately two-thirds remained in experimentation or pilot stages rather than disciplined production deployment [6]. This combination — widespread partial deployment without institutional maturity — is a characteristic condition in which governance failures tend to become systemic, as the survey data reviewed in subsequent sections confirms. Organizations are not uniformly naive about AI risk; they are caught between the organizational momentum of AI adoption and the institutional inertia of governance frameworks that have not kept pace.

The World Economic Forum and Accenture's analysis of responsible AI maturity found that fewer than one percent of organizations have fully operationalized responsible AI practices, with 81 percent still in the earliest maturity stages [7]. Cisco's 2026 Data Privacy Benchmark Study found that while 75 percent of organizations report having dedicated AI governance processes, only 12 percent describe those efforts as mature [8]. The gap between reporting governance and practicing governance is a consistent pattern across every major industry survey of 2025 and early 2026.

This velocity problem is not self-correcting. The commercial pressures that drive AI adoption do not slow as governance matures — they accelerate, because competitive pressure compounds. The organizations that deploy AI fastest accrue early productivity advantages, which creates pressure on competitors to accelerate their own deployments regardless of governance readiness. The result appears to be a market dynamic that devalues investment in governance infrastructure, treating it as friction rather than foundation — a pattern the adoption data throughout this paper is consistent with.

1.2 The Structural Ownership Gap

Governance failures in AI security are not primarily the result of ignorance. Organizations know that AI systems carry risk. The CSA and Google Cloud State of AI Security and Governance survey, conducted across 300 IT and security professionals in late 2025, found that 72 percent of respondents are neutral or lack confidence in their AI security strategy's execution [2]. The 2026 CISO AI Risk Report found that 92 percent of CISOs lack full visibility into AI identities operating in their environments, 86 percent do not enforce access policies for those identities, and 95 percent doubt they could detect or contain misuse of an AI system [4]. These are not gaps in awareness — they are gaps in accountability structures and institutional capacity.

The root cause is structural. AI systems do not map cleanly onto the ownership categories that enterprise security has historically maintained. A traditional enterprise application has an owner (typically a product or IT team), runs on infrastructure with defined security responsibilities (cloud provider, internal operations), and generates outputs that are reviewed by humans before they affect business outcomes. AI systems violate all three of these assumptions. Ownership is disputed between business units, IT, data science teams, and security organizations. Infrastructure responsibilities are blurred by the layered nature of AI supply chains, where a single deployed system may draw on foundation models, fine-tuning pipelines, vector databases, retrieval systems, and integration middleware owned by different internal and external parties. And outputs increasingly affect business outcomes autonomously, without human review gates that allow security failures to be intercepted before they cause harm.

The CSA AI Controls Matrix (AICM) has codified this complexity through its Shared Security Responsibility Model, which defines distinct accountability obligations across five provider roles: AI Customers, Cloud Service Providers, Application Providers, Model Providers, and Orchestrated Service Providers [9]. In practice, organizations deploying AI frequently operate across multiple of these roles simultaneously — building proprietary applications on third-party foundation models, deploying through cloud infrastructure, and orchestrating external service integrations — without clearly mapping security responsibilities to each layer. The result is a patchwork accountability structure that leaves critical control domains unowned.

1.3 Scope of This Analysis

This paper examines four intersecting dimensions of the AI security ownership crisis. The first is the internal governance vacuum — the organizational patterns through which security responsibility for AI becomes unassigned or contested. The second is the shadow AI problem — the deployment of AI tools outside institutional governance structures, which expands the attack surface faster than it can be inventoried or controlled. The third is the agentic inflection point — the specific and severe governance challenges introduced by autonomous AI agents capable of taking independent action in enterprise environments. The fourth is the external pressure context — the convergence of private AI model proliferation, geopolitical volatility, and regulatory acceleration that is simultaneously raising the stakes for governance failures and increasing the difficulty of maintaining governance discipline.

---

2. The Internal Governance Vacuum

2.1 Disputed Ownership and Its Consequences

The most consistent finding across major AI security surveys of 2025 and 2026 is the absence of clear, institutionalized ownership over AI security outcomes. HiddenLayer's 2026 AI Threat Landscape Report, based on a survey of 250 IT and security leaders, found that 73 percent of organizations report internal conflict over who owns AI security controls [1]. This figure is not surprising given the structural dynamics of AI deployment in large organizations: AI projects typically originate in business units or data science teams, infrastructure is provisioned by IT or cloud operations, security review is expected from the CISO organization, and regulatory compliance obligations are managed by legal and risk functions. Each of these groups has a legitimate claim to some aspect of AI security governance, and none has authority over the others.

The HiddenLayer report found that when organizations are asked to identify the primary owner of AI security decisions, CIOs hold that authority in 29 percent of cases, CTOs in approximately 22 percent, and CISOs in 14.5 percent [1]. This distribution reveals a governance structure in which the function most directly responsible for security outcomes — the CISO organization — is operating in a minority position relative to functions whose primary mandate is capability delivery rather than risk management. The predictable consequence is that security requirements are treated as negotiable constraints on deployment timelines rather than non-negotiable prerequisites for deployment authorization.

The Splunk 2026 CISO Report, based on 650 global CISO interviews, documented the operational consequence of this structural mismatch: 96 percent of CISOs have been assigned formal responsibility for AI governance and risk management on top of their existing mandates, yet two-thirds report moderate to significant burnout, and nearly three-quarters are now worried about personal liability for security incidents they cannot prevent because they do not hold decision-making authority over the deployments generating those risks [3]. The CISO mandate has expanded faster than the CISO's institutional authority to fulfill it. This is not only a resource problem — it is fundamentally a governance design problem, in which authority to govern has not kept pace with responsibility to govern.

2.2 The Accountability Inventory Failure

A prerequisite for any governance function is knowing what it governs. In AI security, this basic requirement — maintaining a complete inventory of deployed AI systems — is unmet at the majority of organizations. The Conference Board and ESGAUGE found that while 72 percent of S&P 500 companies disclosed material AI

risks in 2025 annual filings, up from 12 percent in 2023, only a small fraction had formalized processes for inventorying and classifying the AI systems generating those risks [10]. Organizations are disclosing that AI risk exists without having the institutional infrastructure to manage that risk systematically.

The specific consequences of inventory failure compound over time. Without an inventory, risk cannot be classified, controls cannot be assigned, monitoring cannot be scoped, and incidents cannot be attributed to specific systems when they occur. HiddenLayer's 2026 survey found that 31 percent of organizations do not know whether they experienced an AI-related security breach in the preceding twelve months [1]. This is not a failure of detection technology — it is a failure of governance architecture. When AI system ownership is unclear, when AI systems are not registered in any asset management system, and when monitoring is not deployed against AI-specific threat vectors, the absence of evidence of breach is not evidence of absence.

The pattern identified by security architect Ajith Vallath Prabhakar as the "architecture gap" captures six consistent structural failures that sustain this condition: organizations lack complete inventories of their AI systems; ownership of those systems is undefined, causing security incidents to stall at the accountability question; risk classification is inconsistent, with systems of equivalent risk treated differently based on the team that deployed them; monitoring is limited to development environments rather than production; updates are deployed without security review; and security controls are mismatched to actual impact because the impact has not been formally assessed [11]. Each of these failures individually creates exploitable risk. Together they create an institutional condition in which AI security governance cannot function.

2.3 Quantifying the Readiness Gap

The CSA and Google Cloud State of AI Security and Governance survey provided the most granular available picture of organizational readiness across the major domains of AI security governance. 26 percent of organizations have developed comprehensive AI security governance policies [2]. Twenty-seven percent are confident in their ability to secure AI in core business operations [2]. Organizations with comprehensive governance frameworks are two to three times more likely to have adopted agentic AI capabilities, deployed AI-specific monitoring, and expressed confidence in their AI security strategy — suggesting that governance maturity may compound advantage over time, though the correlation may also reflect broader organizational maturity driving both outcomes simultaneously [2].

The maturity gap has measurable financial dimensions. Organizations with mature AI governance platforms are 3.4 times more likely to achieve high governance effectiveness, according to Gartner's Q2 2025 survey of 360 organizations [12]. AI governance platform spending is projected at $492 million in 2026 and is expected to exceed $1 billion by 2030 as regulatory pressure and incident costs drive belated investment [12]. The IBM 2025 Cost of Data Breach Report found that AI-associated breaches carry an average cost exceeding $650,000 above baseline, with shadow AI incidents — where unauthorized AI tools contributed to breach conditions — averaging $4.63 million per event, approximately $670,000 more than standard

breaches [26][13]. The pattern is consistent across multiple analytical frameworks: organizations that underinvest in AI governance pay for it in incident costs, regulatory penalties, and compound risk accumulation that makes future governance harder to achieve.

---

3. Shadow AI and the Ungoverned Enterprise

3.1 The Scope of Unauthorized Deployment

Shadow AI — the deployment and use of AI tools by employees and business units outside institutional oversight — has become the most widespread and least controlled vector of AI security risk in the enterprise. HiddenLayer's 2026 AI Threat Landscape Report found that 76 percent of organizations identify shadow AI as a definite or probable problem, an increase of 15 percentage points over the 61 percent who identified it as a problem in 2025 [1]. The 2026 CISO AI Risk Report found that 75 percent of CISOs have discovered unsanctioned AI tools operating in their environments [4]. Survey data aggregated across multiple 2025-2026 sources suggests that approximately 60 percent of employees use AI tools at work that have not been approved by IT or security, and that 65 percent of AI tools operating in enterprise environments do so without IT oversight [13].

These figures reflect a fundamental mismatch between the governance architecture organizations have built and the behavioral reality of their workforce. AI productivity tools have become consumer-grade in accessibility — individuals can deploy capable AI assistants, code generation tools, document analyzers, and data processing pipelines with no more friction than downloading an application or creating a free account. The governance infrastructure of most enterprises was not designed for a threat model in which employees are simultaneously the most significant source of data exposure risk and the primary drivers of competitive productivity enhancement. The resulting tension — between enabling AI productivity and maintaining security governance — is being resolved by default in favor of productivity, because the productivity benefit accrues immediately to the individual deploying the tool while the security cost is diffuse, delayed, and falls on the organization rather than the individual.

3.2 The Data Exposure Dimension

The core security risk of shadow AI is not adversarial attack against AI systems — it is data exposure through uncontrolled AI inputs. When employees paste sensitive documents into consumer AI assistants, upload proprietary data to unauthorized analysis tools, or build workflows that route customer data through external AI APIs, they are creating data flows that are invisible to the organization's security monitoring, uncontrolled by its data governance policies, and potentially retained, processed, and used by third-party AI providers in ways that violate the organization's legal obligations and customer commitments.

Survey data indicates that 27 percent of organizations report more than 30 percent of their AI-processed data contains private information [13]. This exposure is not distributed uniformly — it concentrates in precisely the functions where AI productivity tools are most compelling: legal, human resources, customer

service, finance, and product development. These are also the functions most likely to handle information protected by sector-specific regulation, contractual obligation, or intellectual property constraints. The McHire incident — in which McDonald's AI-powered hiring platform reportedly exposed approximately 64 million job application records due to default administrative credentials that no security team had reviewed because the system had been deployed without standard security oversight — illustrates the pattern: unauthorized or inadequately governed AI deployments create data exposure at scale before any security function has the opportunity to assess or mitigate the risk [14].

Only 37 percent of organizations have AI governance policies that specifically address employee use of external AI tools, meaning that in 63 percent of organizations, employees are making individual judgments about what data can be shared with which AI systems — judgments made without security training specific to AI risks, without awareness of third-party AI provider data retention and training practices, and without accountability mechanisms that create consequences for exposure [13]. Gartner projects that by 2030, more than 40 percent of enterprises will have experienced a security or compliance incident linked to unauthorized shadow AI deployments made between 2024 and 2028 — incidents that are being seeded now, in the absence of governance [12].

3.3 Shadow AI in the Agentic Era

Shadow AI risk is entering a qualitatively different phase as consumer and prosumer agentic tools become available. The risk profile of an employee using a consumer AI assistant to summarize documents is substantially different from the risk profile of an employee deploying an autonomous AI agent with access to their email, calendar, file storage, and CRM credentials. The former creates data exposure risk. The latter creates credential exposure risk, action execution risk, and supply chain risk — the tools, plugins, and extensions that agentic platforms use may themselves introduce malicious content that executes with the permissions of the deployed agent.

The CSA's analysis of personal AI desktop agents identified what it termed the "Lethal Quartet": an agent configuration in which the agent simultaneously has access to private organizational data, exposure to untrusted external content, capability to communicate externally, and persistent memory across sessions [15]. Each of these capabilities is individually present in most consumer productivity AI tools. Their combination in an agent operating with enterprise credentials creates a condition for catastrophic data exfiltration or organizational manipulation that requires no adversarial attack against the AI system itself — only the successful execution of prompt injection through any of the external content channels the agent processes as part of its normal operation.

The hardening and governance guidance that the CSA has developed for enterprise agent platforms addresses these risks through eight control domains: installation and update security, tool and extension vetting, MCP server governance, runtime sandboxing and isolation, file system and network controls, identity and credential management, behavioral monitoring and anomaly detection, and incident response

[16]. But this guidance presupposes that the agent deployment is known to the security organization and subject to institutional governance. For shadow AI agents deployed by individual employees or business units outside IT oversight, none of these controls are applied — and the organization has no mechanism to detect that the exposure exists.

---

4. The Agentic Inflection Point

4.1 Why Agentic AI Changes the Risk Equation

HiddenLayer's 2026 Threat Landscape Report characterizes the current moment precisely: "AI systems are no longer just generating outputs — they are taking action" [1]. This transition from inference to action is the defining change that makes the governance failures documented in prior sections acutely dangerous rather than merely operationally expensive. An AI system that generates incorrect or malicious content presents a risk that a human reviewer may catch. An AI system that autonomously executes actions — sending communications, making API calls, modifying files, delegating tasks to other agents, and triggering downstream workflows — presents a risk that may materialize before any human has the opportunity to intervene.

Industry projections indicate that 40 percent of enterprise applications will feature task-specific AI agents by the end of 2026, up from fewer than 5 percent in 2025 [12]. Industry analysts project that approximately one-third of enterprise software applications may incorporate agentic AI by 2028 [17]. These agents do not operate in isolation; they interact with enterprise systems through credentials, access APIs through authenticated connections, and increasingly delegate tasks to other agents through multi-hop orchestration chains. The security infrastructure of most enterprises was not designed for a population of autonomous, privileged, machine-identity-bearing actors whose deployment is accelerating faster than IAM governance can absorb.

4.2 The Non-Human Identity Crisis

The most immediate and measurable dimension of the agentic governance problem is the non-human identity (NHI) crisis. AI agents are identities — they require credentials to access databases, cloud services, code repositories, communication platforms, and other systems. As agent deployment accelerates, the NHI population is growing faster than organizational IAM infrastructure can govern it. Identity security research indicates that NHIs already outnumber human identities in the average enterprise environment at a ratio of 50 to 1, with projections suggesting this ratio may reach 80 to 1 within two years [27]. The CSA's Agentic Identity Survey found that 78 percent of organizations lack documented policies for creating or removing AI identities, 92 percent are not confident their legacy IAM approaches can manage AI identity risks, and 28 percent can trace agent actions back to a human sponsor across all environments [25].

The CSA Securing the Agentic Control Plane whitepaper has identified the critical failure mode: agent identities "often begin expanding before security teams can define how to control them" and frequently operate "with no clear owner and outside traditional enforcement models" [19]. When agents delegate tasks

to sub-agents, credentials compound: the delegating agent must pass authority to the sub-agent, but conventional IAM systems were not designed for ephemeral, machine-generated delegation chains. Each delegation hop in a multi-agent pipeline introduces a trust boundary where the receiving agent must be authenticated, its permissions scoped, and its actions logged — requirements that most current agent frameworks fulfill poorly, if at all.

The Cybersecurity Insiders 2026 CISO AI Risk Report provides the operational view of this failure: 71 percent of organizations have AI tools with access to core business systems including ERP and CRM platforms, but 16 percent report applying formal governance controls to that access [4]. Eighty-six percent lack or do not enforce access policies for AI identities [4]. 19 percent govern even half of their generative AI accounts with the same discipline they apply to human user accounts [4]. And just 5 percent of security leaders feel prepared to detect and contain a compromised AI agent [4]. The McKinsey "Lilli" internal AI platform breach during a red team exercise, in which the platform was compromised in under two hours, illustrates the operational consequence of this readiness gap: organizations are deploying agents with privileged access to sensitive systems into environments where the security team's capability to detect, respond to, and contain an agent compromise is essentially theoretical [13].

4.3 Agentic Attack Vectors

The agentic threat landscape is rapidly developing its own distinct attack taxonomy, documented by MITRE ATLAS, the OWASP Top 10 for Agentic Applications, and the CSA MAESTRO framework for multi-agent system threat modeling. Understanding the specific attack vectors that the governance ownership crisis enables is essential to understanding why structural remediation — not incremental control improvement — is the appropriate response.

The most consequential attack vector is prompt injection through agent input channels. As HiddenLayer's principal security researcher Marta Janus has noted, "as agents can browse the web, execute code, and trigger workflows, prompt injection becomes an operational security risk" rather than merely a quality concern [1]. Unlike traditional prompt injection in conversational AI, which produces malicious outputs, agentic prompt injection through web browsing, document processing, email reading, or tool result parsing produces malicious actions — actions executed autonomously by the agent, with its full credential set, before any human reviewer has the opportunity to intervene.

The Model Context Protocol (MCP), which has emerged as the dominant standard for agent-to-tool communication, has accumulated a substantial vulnerability record. Between January and February 2026, security researchers filed over 30 CVEs targeting MCP servers, clients, and infrastructure [19]. Critical vulnerabilities include pre-authentication remote code execution in MCP client implementations, tool poisoning through hidden malicious instructions embedded in tool descriptions, rug pull attacks in which tool definitions are modified after approval to introduce malicious behaviors, and session hijacking through predictable session identifiers. The OWASP Top 10 for Agentic Applications, published in December 2025,

identified agent goal hijacking, tool misuse and exploitation, identity and privilege abuse, and insecure inter-agent communication as the four highest-priority risk categories — each of which is directly enabled by governance failures that leave agent identities unscoped, tool approvals ungoverned, and agent behaviors unmonitored [24].

The threat surface extends through the agentic supply chain. The CSA's research on the OpenClaw platform documented a malicious campaign — ClawHavoc — through which 1,467 malicious skills were published to the community skill repository, with 91 percent combining prompt injection payloads with malware delivery [16]. This attack succeeded not because the platform lacked security controls but because no institutional process existed for vetting third-party capabilities before deployment — the same governance gap that allows shadow AI tools to enter enterprise environments without review.

4.4 Governance Requirements for Agentic Systems

The CSA Capabilities-Based Risk Assessment (CBRA) framework, developed in 2025, provides a structured approach to risk-proportional governance for AI systems through a multiplicative risk scoring model: System Risk = Criticality × Autonomy × Permission × Impact [20]. For agentic systems, each of these dimensions is elevated relative to conventional AI deployments. Autonomy levels are high by design. Permission scopes expand as agents connect to more enterprise systems. Impact radius extends through action chaining and delegation. The CBRA model generates high or very high risk scores for most production agent deployments, indicating that the corresponding control tier — comprehensive AICM control coverage, behavioral monitoring, human oversight gates for consequential actions, and regular re-certification — is not optional for responsible deployment.

The minimum viable control stack for agentic systems, as detailed in CSA's Securing the Agentic Control Plane framework, includes six elements: unique, cryptographically verifiable agent identity; short-lived, scoped credentials that do not persist beyond the task requiring them; policy enforcement gates that evaluate every tool call against authorization policy before execution; sandboxing that constrains the blast radius of agent compromise; approval workflows that route consequential actions through human review; and complete action lineage that provides end-to-end attribution for every agent action back to the authorizing human instruction [19]. Organizations deploying agents without this control stack are not accepting elevated risk within a governance framework — they are operating outside any governance framework, with the attendant exposure to both security incidents and the personal liability implications that the Splunk CISO survey has documented.

---

5. External Pressures Compounding Structural Risk

5.1 The Private AI Model Explosion

Forrester's analysis, published in March 2026, identifies a fundamental shift in enterprise AI architecture that will increase the governance burden over the coming years: the movement toward proprietary, private AI models deployed on private infrastructure, driven by enterprise concerns about data confidentiality and model trust [5]. Forrester projects that 70 percent of AI-generated enterprise revenue within five years will derive from private models rather than public ones, with at least 15 percent of enterprises beginning significant private model deployments in 2026 [5].

This shift has a specific security implication. Public AI models from major providers — despite their limitations — are deployed through managed APIs that include at least some degree of vendor-side security monitoring, abuse detection, and incident response capability. Private model deployments transfer security responsibility entirely to the deploying organization. The discipline that applies to public model integration — API key management, rate limiting, output monitoring — must be substantially extended for private model infrastructure, which transfers full security responsibility, including training data protection, model access controls, inference infrastructure hardening, and fine-tuning pipeline security, entirely to the deploying organization. Organizations that are already struggling to govern their use of third-party AI models will face substantially greater challenges governing the private AI infrastructure they are planning to build.

The governance readiness data does not suggest that organizations are prepared for this transition. Twenty-five percent of planned AI enterprise spend is expected to be deferred from 2026 to 2027 as financial rigor slows production deployments [5]. This deferral is not being driven by security governance maturity — it is being driven by cost discipline and ROI concerns. The security governance gap is not a reason for deferral in the current framing; it is a risk that is being accepted by default.

5.2 Geopolitical Volatility and the CISO Mandate

Forrester analyst Stephanie Balaouras, writing in March 2026, identified geopolitical volatility as an emerging technology leadership test that demands sharper trade-offs, stronger resilience, and faster decisions from CIOs and CISOs [21]. The convergence of geopolitical pressure with the AI security ownership crisis creates a specific compounding risk: CISOs who are already overextended by the expansion of AI governance responsibilities now face concurrent pressure to accelerate AI adoption for competitive and operational resilience reasons, tighten controls against state-sponsored threats, and maintain compliance with an expanding regulatory landscape.

The regulatory dimension is particularly acute. The EU AI Act's obligations for high-risk AI systems — human oversight requirements, continuous risk management, cybersecurity resilience, and automatic event logging — become enforceable in August 2026, with penalties up to €15 million or 3 percent of global annual revenue for high-risk system violations, with the highest tier reaching €35 million or 7 percent for prohibited AI practices [18]. For organizations with EU operations that have not achieved compliance readiness, this creates a hard deadline that will collide with the broader AI governance maturity gap documented throughout this paper. The Conference Board found that 72 percent of S&P 500 companies disclosed material AI risks in their 2025 annual filings, up from 12 percent in 2023, reflecting awareness that regulatory and liability exposure is increasing — but disclosure without remediation is not a governance strategy [10].

The geopolitical dimension also creates direct threat amplification for organizations with inadequate AI governance. State-sponsored threat actors have demonstrated capability to leverage AI systems as orchestration tools for cyber operations. Threat intelligence from 2025 documented a case in which a foundation model was used as an orchestrator for an espionage campaign against 30 organizations, automating reconnaissance and tool-chaining at a scale and speed that would have been prohibitive with conventional attack infrastructure [14]. Organizations that have deployed AI agents with broad access to enterprise systems, without adequate behavioral monitoring or containment capability, present attractive targets for this attack pattern: compromise of an agent with broad permissions is equivalent to credential compromise at scale, but with automated execution capability that dramatically compresses the attack timeline.

---

6. A Framework for Structural Remediation

6.1 Governance Architecture Principles

Addressing the AI security ownership crisis requires structural interventions at the organizational, process, and technical layers. Incremental improvements to individual controls will not resolve a failure rooted in the absence of clear accountability, comprehensive inventory, and institutionalized governance processes. The remediation framework presented here is organized around five principles drawn from CSA guidance, industry best practice, and the documented patterns of governance failure analyzed in prior sections.

The first principle is that ownership must be explicit, assigned, and enforced. Every AI system deployed in the enterprise must have a named owner — an individual or team responsible for its security governance — and that ownership must be documented in an asset registry, reviewed periodically, and updated when systems change or ownership transfers. The second principle is that governance must be proportional to risk. The CSA CBRA framework's multiplicative risk scoring model provides a practical mechanism for differentiating governance intensity based on system autonomy, permission scope, criticality, and impact radius — avoiding both governance gaps in high-risk systems and governance overhead in low-risk applications. The third principle is that shadow AI must be addressed through both detection and channeling: detection to identify unauthorized deployments and assess their risk, and channeling to provide sanctioned alternatives that meet employees' legitimate productivity needs without creating ungoverned exposure. The fourth principle is that agent identity must be governed with the same rigor as human identity. The fifth principle is that governance must generate observable evidence — audit trails, behavioral telemetry, and incident records — that allow governance effectiveness to be assessed and improved over time.

6.2 Organizational Accountability Structure

The most critical structural intervention is establishing a clear, institutionalized accountability model for AI security that resolves the ownership conflicts documented throughout this paper. The CSA AI Organizational Responsibilities framework, developed through the AICM working group, provides a detailed RACI model for AI security accountability that defines specific responsibilities for each actor in the AI deployment ecosystem — executive leadership, the CISO organization, business unit owners, data science teams, IT operations, legal and compliance, and HR [22].

For organizations addressing the ownership vacuum, the practical implementation of this framework requires three elements. First, formal designation of an AI Security Lead — a function with clear authority over AI security policy, governance standards, and deployment authorization, regardless of where those

deployments originate. Second, an AI Governance Review Board that provides cross-functional oversight for significant AI deployments, with mandatory participation from the CISO organization and a charter that gives security review a blocking role in deployment authorization. Third, an AI Asset Registry that captures the minimum governance metadata for every AI deployment: the owning team, the business function served, the data it accesses, the permissions it holds, the monitoring coverage applied, and the last security review date. Organizations that implement these three structural elements will have the institutional foundation necessary to apply more specific technical controls systematically.

The following table summarizes the primary governance gaps identified in this analysis and the CSA framework controls that address each:

Governance Gap	Risk Description	Primary AICM Control Domain	Supporting Framework
Undefined AI ownership	No clear accountability for security outcomes	Governance, Risk and Compliance	CSA AI Organizational Responsibilities
Incomplete AI inventory	Unknown deployment scope, unmonitorable	Change Control and Configuration Management	CBRA risk scoring
Shadow AI deployment	Ungoverned data exposure and agent risk	Data Security and Privacy	Zero Trust; Shadow Access guidance
Unscoped agent permissions	Excessive privilege enabling broad breach impact	Identity and Access Management	MAESTRO; OWASP ASI03
No behavioral monitoring	Compromised agents operate undetected	Logging and Monitoring	MAESTRO Layer 5; MITRE ATLAS
Ungoverned NHI lifecycle	Agent credentials persist beyond authorization	Identity and Access Management	CSA Agentic IAM framework
No incident response for AI	Inability to contain AI-enabled breaches	Security Incident Management	CBRA circuit breaker guidance
Supply chain vetting gaps	Malicious tools and models enter production	Supply Chain Management	AICM SCM controls; OWASP ASI04

6.3 Shadow AI Governance

Effective shadow AI governance requires abandoning the assumption that prohibition is a viable strategy. Approximately 60 percent of employees who use unauthorized AI tools do so because those tools provide genuine productivity benefits that sanctioned alternatives do not yet deliver [13]. A governance approach that attempts to block all unauthorized AI use without providing sanctioned alternatives will fail — employees will route around the restriction, security teams will spend resources on enforcement rather than risk management, and the governance friction will create organizational resistance that undermines the broader AI security program.

The CSA Shadow Access and AI framework recommends a four-stage lifecycle approach: continuous monitoring to detect unauthorized AI deployments as they emerge; context and visualization to understand what data these deployments access and what risk they represent; automated risk analysis using CBRA-aligned scoring to prioritize governance response; and remediation through a combination of sanctioned alternatives, policy controls, and targeted enforcement for the highest-risk deployments [23]. This approach is materially different from a blocking posture: it is designed to maintain visibility across the full shadow AI population, concentrate governance effort on the deployments presenting meaningful risk, and provide organizational leadership with a credible picture of AI exposure that enables informed risk acceptance decisions.

The AICM Governance, Risk and Compliance domain provides the policy framework for shadow AI governance, requiring documented acceptable use policies, risk assessment processes for new AI tool adoption, and employee training on AI security risks [9]. The Identity and Access Management domain controls the access credentials that enable shadow AI tools to connect to enterprise systems — implementing credential brokering, just-in-time access, and scope limitations can materially constrain the blast radius of shadow AI deployments even when they cannot be fully prevented [9].

6.4 Agentic Governance Implementation

For organizations deploying or planning to deploy autonomous agents, the governance requirements are more demanding than for conventional AI systems, and the timeline for implementation is compressed by the pace of agentic AI adoption. The CSA Securing the Agentic Control Plane framework provides a phased implementation roadmap organized around the six strategic programs of the CSAI Foundation: an AI Risk Observatory for real-time visibility into agent behavior and vulnerabilities; Agentic Best Practices defining the security controls and development lifecycle requirements for secure agent deployment; Education and Credentialing programs building workforce readiness; CxOtrust executive engagement translating agentic risk into board-level decisions; Global Assurance and Trust through STAR for AI certification extensions; and Future Forward Initiatives addressing agent certification and catastrophic risk research [19].

For organizations at early maturity stages, the immediate priority is establishing the identity governance foundation: every deployed agent must have a registered, cryptographically verifiable identity; credential scopes must be defined as narrowly as the agent's task requires; and a human sponsor must be identified for every agent identity in the registry. The MAESTRO threat modeling framework provides a seven-layer structure for analyzing multi-agent system risks — from the foundation model layer through orchestration, memory, data access, and external interface layers — that allows security teams to identify the specific threat vectors relevant to their deployment architecture and prioritize controls accordingly [9].

Short-term implementation (three to six months) should focus on behavioral monitoring coverage for all production agents, implementation of policy gates on consequential tool calls, and establishment of approval workflows for agent actions that exceed defined risk thresholds. The CBRA circuit breaker concept — predefined conditions under which an agent is automatically suspended and queued for human review — provides a practical mechanism for containing agent risk without requiring full-time human oversight of every agent action [20]. Medium-term implementation (six to twelve months) should extend to supply chain governance for agent tools and extensions, Zero Trust network segmentation of agent communication, and formal alignment with NIST AI RMF governance functions. Organizations with EU operations must treat the August 2026 EU AI Act enforcement date as a hard deadline for implementing human oversight, risk management, and logging controls for any AI systems that meet the high-risk classification.

---

7. Conclusions and Recommendations

7.1 The Compounding Cost of Inaction

The AI security ownership crisis is not a static condition — it is a compounding one. Every AI deployment that enters production without clear ownership, defined controls, and behavioral monitoring expands the ungoverned surface area that security teams cannot see, cannot manage, and cannot respond to when it generates an incident. Every agent identity created without an IAM policy, every shadow AI tool deployed without data governance, and every agentic workflow executed without action lineage adds to an accumulating liability that will eventually materialize as either a security incident, a regulatory penalty, or both.

Forrester has explicitly predicted a publicly disclosed breach attributable to an agentic AI deployment in 2026 [5], and HiddenLayer found that one in eight reported AI-related breaches is already linked to agentic systems [1]. The EU AI Act enforcement date arrives in August 2026 for organizations that have not yet achieved compliance readiness. And the private AI model explosion documented by Forrester will transfer even more security responsibility to organizations that are not yet equipped to manage the governance requirements of the models they already have [5].

7.2 Recommendations

The following recommendations are organized by time horizon and draw directly on CSA framework guidance:

Immediate Actions (0 to 90 days)

Organizations should begin with a governance baseline assessment using the CSA and Google Cloud State of AI Security maturity benchmarks and the AICM AI-CAIQ self-assessment questionnaire to establish a documented understanding of current governance posture [2][9]. This assessment should produce a preliminary AI asset inventory, an ownership assignment for each identified system, and a prioritized list of governance gaps organized by risk. The CBRA risk scoring model provides the analytical framework for risk prioritization [20]. Simultaneously, shadow AI discovery should be initiated through network monitoring, data loss prevention tooling, and employee survey to identify the population of unauthorized AI tools in use and the data flows they create.

Short-Term Mitigations (90 days to 6 months)

Governance structure formalization should establish the AI Security Lead function, the AI Governance Review Board, and the AI Asset Registry described in Section 6.2. Shadow AI governance should transition from discovery to the four-stage lifecycle approach recommended by the CSA Shadow Access framework [23]. For any agentic AI deployments in production or active pilot, immediate implementation of agent identity registration and credential scoping should be prioritized, followed by behavioral monitoring deployment and policy gate implementation on high-risk tool calls. Organizations approaching the August 2026 EU AI Act enforcement date for high-risk AI system obligations should initiate a regulatory gap assessment against the specific requirements of Articles 9, 12, 14, and 15.

Strategic Considerations (6 to 18 months)

Full AICM alignment across the 18 control domains, with controls deployed proportionally to CBRA risk scores, should be the medium-term governance target. MAESTRO threat modeling should be applied to all multi-agent deployments, with threat model outputs informing monitoring configuration and incident response planning. STAR for AI Level 1 self-assessment should be completed and submitted to establish a transparency baseline and identify gaps requiring remediation before pursued certifications. For organizations with significant private AI investments planned, governance infrastructure for private model security — training data protection, inference access control, fine-tuning pipeline security, and model supply chain vetting — should be developed in advance of model deployment, not after.

CSA Resource Alignment

The analysis and recommendations presented in this paper draw directly on the following CSA frameworks, publications, and initiatives:

AI Controls Matrix (AICM) v1.0 provides the foundational control framework for AI security governance, with 243 control objectives across 18 domains covering all major governance gaps identified in this paper. The AICM's Shared Security Responsibility Model directly addresses the ownership clarity problem by defining accountability across five provider role types.

Capabilities-Based Risk Assessment (CBRA) provides the risk-proportional governance methodology recommended for prioritizing AI asset governance and differentiating control intensity based on system autonomy, permission scope, criticality, and impact.

MAESTRO provides the multi-agent system threat modeling framework referenced in Section 4 for analyzing agentic-specific attack vectors across seven system layers.

AI Organizational Responsibilities (Parts 1 and 2) provides the RACI model for AI security accountability and the policy framework for shadow AI governance.

Shadow Access and AI and **Confronting Shadow Access Risks** provide the governance lifecycle and Zero Trust integration guidance for managing shadow AI deployments.

Securing the Agentic Control Plane provides the comprehensive strategic framework for agentic AI governance, including the CSAI Foundation's six strategic programs and the phased implementation roadmap referenced in Section 6.4.

Agentic AI Identity and Access Management provides the technical framework for managing non-human identities using Decentralized Identifiers, Verifiable Credentials, and Zero Trust principles.

STAR for AI Program provides the assurance and certification pathway through which organizations can demonstrate AI security governance maturity to external stakeholders and regulators.

State of AI Security and Governance 2025 (CSA/Google Cloud) provides the benchmark maturity data cited throughout this paper for contextualizing organizational governance posture.

References

[1] HiddenLayer, "2026 AI Threat Landscape Report," HiddenLayer.com, March 18, 2026. https://www.hiddenlayer.com/news/hiddenlayer-releases-the-2026-ai-threat-landscape-report-spotlighting-the-rise-of-agentic-ai-and-the-expanding-attack-surface-of-autonomous-systems

[2] Cloud Security Alliance and Google Cloud, "State of AI Security and Governance Report 2025," Cloud Security Alliance, December 2025. https://cloudsecurityalliance.org/artifacts/the-state-of-ai-security-and-governance

[3] Splunk, "2026 CISO Report: Agentic AI and Digital Resilience," Cisco/Splunk, February 2026. https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2026/m02/splunk-report-agentic-ai-takes-center-stage-in-cisos-path-to-digital-resilience.html

[4] Cybersecurity Insiders, "2026 CISO AI Risk Report," Cybersecurity Insiders, 2026. https://www.cybersecurity-insiders.com/2026-ciso-ai-risk-report/

[5] Forrester, "The Private AI Model Explosion," Forrester Blogs, March 19, 2026. https://www.forrester.com/blogs/the-private-ai-model-explosion/

[6] McKinsey & Company, "The State of AI in 2025," McKinsey Global Survey, 2025. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[7] World Economic Forum and Accenture, "Advancing Responsible AI Innovation: A Playbook," World Economic Forum, September 2025. https://www.weforum.org/publications/advancing-responsible-ai-innovation-a-playbook/

[8] Cisco, "2026 Data Privacy Benchmark Study," Cisco Systems, 2026. https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2026/m01/ai-data-privacy-investments-governance-cisco-report.html

[9] Cloud Security Alliance, "AI Controls Matrix (AICM) v1.0 and Implementation Guidance," Cloud Security Alliance, July 2025. https://cloudsecurityalliance.org/artifacts/ai-controls-matrix

[10] The Conference Board and ESGAUGE, "AI Risk Disclosure in S&P 500 Annual Reports," October 2025. Referenced via SecurePrivacy.ai AI Risk and Compliance 2026. https://secureprivacy.ai/blog/ai-risk-compliance-2026

[11] Ajith Vallath Prabhakar, "The Architecture Gap: Why Enterprise AI Governance Fails," December 14, 2025. https://ajithp.com/2025/12/14/enterprise-ai-governance-framework/

[12] Gartner, "AI Governance Platform Market Forecast and Adoption Survey," Gartner Research, February 2026 and Q2 2025. Referenced via Dataversity AI Governance in 2026. https://www.dataversity.net/articles/ai-governance-in-2026-is-your-organization-ready/

[13] Bessemer Venture Partners, "Securing AI Agents: The Defining Cybersecurity Challenge of 2026," Bessemer Venture Partners Atlas, 2026. https://www.bvp.com/atlas/securing-ai-agents-the-defining-cybersecurity-challenge-of-2026

[14] ISACA, "Avoiding AI Pitfalls in 2026: Lessons Learned from Top 2025 Incidents," ISACA Now Blog, 2025. https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2025/avoiding-ai-pitfalls-in-2026-lessons-learned-from-top-2025-incidents

[15] Cloud Security Alliance, "Policy on Personal AI Desktop Agents," CSA AI Safety Initiative, 2024. https://cloudsecurityalliance.org/artifacts/policy-on-personal-ai-desktop-agents

[16] Cloud Security Alliance AI Safety Initiative, "Hardening OpenClaw: Enterprise Security Guide for Autonomous Desktop Agents," CSAI v1.0, 2026. https://cloudsecurityalliance.org

[17] Help Net Security, "Non-Human Identities and the AI Identity Explosion," 2025-2026. Referenced via BVP Atlas and Biometric Update RSAC 2026 coverage. https://www.biometricupdate.com/202603/ai-agent-identity-and-next-gen-enterprise-authentication-prominent-at-rsac-2026

[18] European Parliament, "EU Artificial Intelligence Act," Official Journal of the European Union, Regulation (EU) 2024/1689, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[19] Cloud Security Alliance AI Safety Initiative, "Securing the Agentic Control Plane: A Comprehensive Framework for Governing Autonomous AI Systems," CSAI v1.0, March 2026. https://cloudsecurityalliance.org

[20] Cloud Security Alliance, "Capabilities-Based Risk Assessment (CBRA) for AI Systems," CSA, 2025. https://cloudsecurityalliance.org/artifacts/capabilities-based-risk-assessment-for-ai-systems

[21] Stephanie Balaouras, Forrester, "Geopolitical Volatility Has Become a Technology Leadership Test," Forrester Blogs, March 24, 2026. https://www.forrester.com/blogs/

[22] Cloud Security Alliance, "AI Organizational Responsibilities: Governance, Risk Management, Compliance and Cultural Aspects," CSA Best Practices Series, 2024. https://cloudsecurityalliance.org/artifacts/ai-organizational-responsibilities

[23] Cloud Security Alliance IAM Working Group, "Shadow Access and AI," CSA Research Report, 2024. https://cloudsecurityalliance.org/artifacts/shadow-access-and-ai

[24] OWASP Foundation, "OWASP Top 10 for Agentic Applications 2025," December 10, 2025. https://owasp.org/www-project-top-10-for-large-language-model-applications/

[25] Cloud Security Alliance, "Agentic AI Identity and Access Management: A New Approach," CSA AI Safety Initiative, 2025. https://cloudsecurityalliance.org/artifacts/agentic-ai-identity-and-access-management-a-new-approach

[26] IBM Security, "Cost of a Data Breach Report 2025," IBM, 2025. https://www.ibm.com/reports/data-breach

[27] CyberArk, "Identity Security Threat Landscape Report 2025," CyberArk, 2025. https://www.cyberark.com/resources/threat-research-blog/cyberark-identity-security-threat-landscape-2025