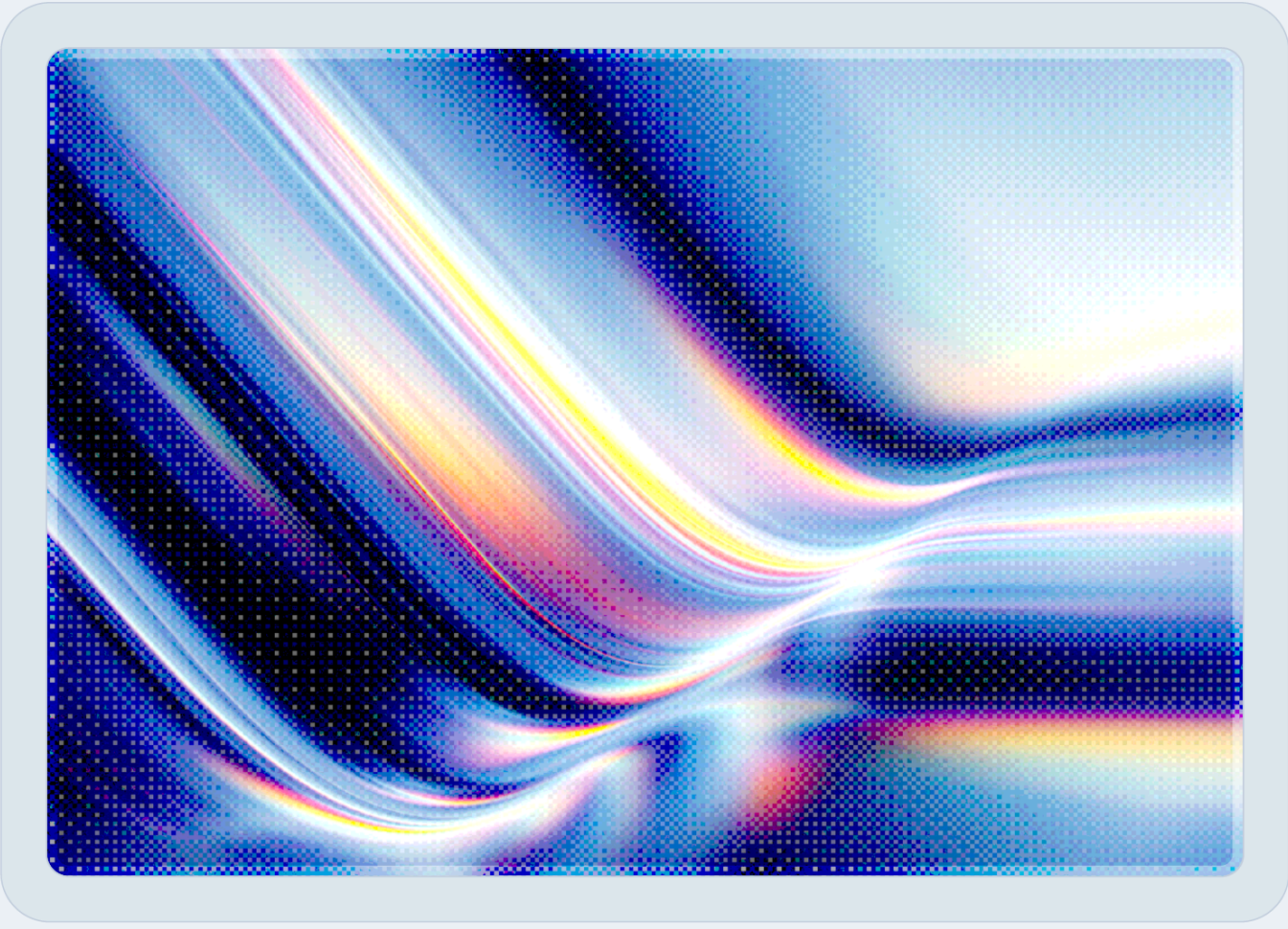
CSAI Foundation | Cloud Security Alliance

The Shadow AI Blind Spot: Ownership Fragmentation as Enterprise Attack Surface

How Ungoverned AI Deployment Fragments Accountability and Creates Systemic Security Risk

2026-05-08

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary

Enterprise security teams have spent decades learning to govern what employees install on managed endpoints, what SaaS platforms procurement approves, and what infrastructure engineers deploy to the cloud. Shadow AI represents a qualitatively different challenge. It is not simply unauthorized software — it is unauthorized cognition embedded in workflows, often by the most technically sophisticated employees in the organization, operating under API keys drawn from legitimate corporate credentials, connecting to data sources that no policy was written to protect.

The evidence of failure is already visible in breach data. Organizations that use high levels of shadow AI experience breach costs an average of $670,000 higher than those with minimal unauthorized AI deployment, according to IBM's 2025 Cost of a Data Breach Report [1]. One in five studied organizations in that same report had experienced a breach linked to shadow AI. More disturbing still is the agentic dimension: in 2026, 65% of organizations report having experienced an AI agent security incident in the past year, and every single organization reporting an incident reported real business impact [2]. The threat is no longer theoretical.

The analysis in this paper argues that ownership fragmentation is the structural factor most amenable to enterprise intervention, and that addressing it is necessary — though not alone sufficient — to reduce the risks that shadow AI creates. Shadow AI is not primarily a training problem, a policy problem, or a technology problem, though all of these contribute. It is an accountability problem. When no function owns the complete lifecycle of an enterprise AI deployment — from provisioning through monitoring through retirement — every phase of that lifecycle becomes an attack surface. Evidence from shadow IT research and enterprise surveys consistently suggests that the employee who installs an unauthorized tool does so in part because the authorized path is unclear, unresponsive, or nonexistent. The orphaned agent that continues running after its creator leaves the organization does so because no governance system tracks agent provenance or mandates retirement. The multi-vendor AI stack that no one can fully audit exists because shared responsibility models designed for cloud infrastructure have not been coherently extended to AI systems.

This paper maps the five structural dimensions of AI ownership fragmentation, analyzes how each creates exploitable attack surface, examines realized breaches and incidents, and presents a prescriptive framework for establishing accountability across the full AI lifecycle. The framework draws on CSA's AI Controls Matrix (AICM), the MAESTRO agentic threat modeling methodology, and the CSA AI Organizational Responsibilities guidance series to provide a practical path for enterprises that need to govern what they have already deployed, not only what they plan to deploy next.

Introduction: The Governance Gap in Plain Sight

The challenge of shadow AI did not emerge suddenly. It is the predictable consequence of a structural gap between the pace of AI adoption and the pace of governance development, a gap that has been widening since generative AI tools became broadly accessible in 2022 and has accelerated dramatically with the emergence of agentic AI frameworks in 2024 and 2025.

What makes the current moment distinct from earlier cycles of ungoverned technology adoption — the personal smartphone wave, the early SaaS proliferation, the initial public cloud rush — is the agentic dimension. Earlier waves introduced data exposure risks when employees used personal accounts or unapproved tools. They introduced shadow IT inventories that were difficult to discover and difficult to remove. They created data governance problems when sensitive information flowed to third-party systems without contractual protections. All of these remain true for shadow AI, but shadow agents add a different category of risk: autonomous action. An ungoverned AI agent does not merely receive data and return a result; it may initiate API calls, modify documents, interact with downstream systems, and take consequential actions in the enterprise environment — all without any human review, without audit trail integrity, and without the accountability chain that responsible AI deployment requires.

The organizational response to shadow AI has largely followed a familiar playbook: issue a policy, provide a list of approved tools, train employees, and apply network-based detection to identify unauthorized AI tool usage. This playbook is necessary but insufficient because it treats shadow AI as a user behavior problem rather than a structural accountability problem. The reason shadow AI proliferates is not primarily that employees are reckless. It is that enterprises have not yet built the institutional structures — the ownership models, the lifecycle governance programs, the cross-functional accountability frameworks — that would make governed AI adoption the path of least resistance.

This paper argues that closing the shadow AI governance gap requires addressing its structural causes, not its symptoms. The following sections examine the anatomy of ownership fragmentation, the attack surface it creates, and the framework elements required to close it.

Defining Shadow AI in the Modern Enterprise

Shadow AI encompasses any AI system, tool, agent, or workflow operating within enterprise scope without the knowledge, authorization, or governance oversight of the security and IT functions responsible for managing enterprise risk. This definition deliberately includes three categories that governance programs often treat separately but that share a common accountability failure.

The first category is unsanctioned tool use: employees accessing publicly available AI services — conversational AI systems, code generation tools, document summarization services — through personal accounts, personal devices, or even corporate devices using mechanisms that bypass enterprise monitoring. The scale of this category is substantial, though estimates vary significantly across studies and survey methodologies. Some analyses find that 68% to 80% of employees use AI tools without IT approval, and that a majority have submitted sensitive company information in the process [3][4]. A 2026 JumpCloud analysis places the figure closer to half of all employees, and notes that enterprise leaders — not frontline staff — are among the most frequent users of unapproved AI tools [5]. The variance likely reflects differences in how "unsanctioned use" is defined and how surveys are administered, but the consistent finding across all methodologies is that the behavior is pervasive and management is not exempt.

The second category is shadow AI infrastructure: AI-powered services deployed by business units, product teams, or individual developers into enterprise environments — often in cloud accounts, SaaS platforms, or development environments — without going through security review, IT procurement, or risk assessment processes. This category includes AI capabilities embedded in approved SaaS platforms that employees activate without IT's knowledge, AI integrations built into approved cloud infrastructure without separate security evaluation, and AI services accessed through shadow cloud accounts that IT has not inventoried. The average enterprise now runs approximately 66 generative AI applications, with an estimated 10% classified as high-risk under emerging governance frameworks [6]. The probability that all 66 applications in any given enterprise went through a rigorous security review is, by the evidence, low.

The third and most operationally significant category is shadow AI agents: autonomous AI workflows and agent deployments that operate in enterprise environments without formal governance oversight. These agents are increasingly deployed not by rogue employees but by technically sophisticated staff — software engineers, data scientists, DevOps practitioners — who are using legitimate tools (LangChain, AutoGPT, CrewAI, the Model Context Protocol) to automate workflows they genuinely believe serve the enterprise's interests. They are authenticated with real corporate API keys and real service account credentials, connecting to real production data sources. In the past year, 82% of organizations have discovered at least one AI agent or autonomous workflow that security or IT did not previously know about [2]. The most common deployment environments for these shadow agents are internal automation systems (51%) and LLM platform integrations (47%) [2].

What these three categories have in common is not employee malice. It is the absence of a clear, usable, responsive governance path. Where shadow AI flourishes, it is almost always in the space between where governance was designed and where AI adoption has actually gone.

Five Dimensions of Ownership Fragmentation

Shadow AI is a symptom. Ownership fragmentation is the disease. The governance failures that produce shadow AI environments stem from five distinct structural dimensions, each of which deserves analysis on its own terms.

The C-Suite Accountability Vacuum

Enterprise AI governance operates in a leadership environment defined by contested ownership. Surveys of large enterprises consistently find that no single executive function has a clear, uncontested mandate for AI security governance. A 2025 Acuvity study found that CIOs control AI security decisions in 29% of organizations, while CISOs rank fourth at just 14.5% [7]. Security teams claim ownership in 39% of organizations, IT departments in 32%, and emerging AI security functions in 13%, with no single function holding a clear majority [7]. Industry observers project that the majority of large enterprises will designate AI leaders in the current governance environment, while cautioning that "title inflation" often masks a lack of real budget or cross-functional authority [8]. The 2026 CISO AI Risk Report found that 96% of CISOs now report responsibility for AI governance and risk management [9], yet the same CISOs also report that many of the AI systems generating risk were deployed without their knowledge or approval.

This is not a disagreement about strategy. It is a structural gap. When the CISO is accountable for AI security risk but does not own the procurement process, the cloud environment, or the software development lifecycle where AI capabilities are being introduced, accountability without authority produces exactly what we observe: awareness of the problem combined with limited capacity to address it. When the CIO owns technology procurement but lacks the mandate and expertise to evaluate AI-specific security risks, approvals happen based on general technology criteria rather than the threat-specific assessments that AI systems require. When product and engineering leaders introduce AI capabilities into their platforms in the course of normal development without a governance trigger requiring them to seek security review, the accountability gap becomes self-perpetuating.

The emerging organizational response — the Chief AI Officer (CAIO) role, the AI governance board, the cross-functional AI risk committee — represents a genuine attempt to address this fragmentation. But the effectiveness of these structures depends entirely on whether they have clear decision rights, funded mandates, and the cross-functional authority to govern what has already been deployed, not only what is being evaluated for future deployment.

Shadow AI versus Shadow IT: A Qualitatively Different Attack Surface

Shadow IT was primarily a data exposure and compliance problem. Employees using personal Dropbox accounts, unapproved collaboration tools, or unauthorized project management software created risks around data residency, contractual protections, and regulatory compliance. The attack surface was predominantly about what data went where. Shadow AI introduces a fundamentally different attack surface because AI systems do not merely store and transmit data — they process it, act on it, generate outputs that may be used in consequential decisions, and, in the agentic case, take autonomous actions in the enterprise environment.

Consider the difference between an employee using a personal file-sharing service and an employee using an unauthorized AI coding assistant. The file-sharing risk is data exfiltration. The coding assistant risk includes data exfiltration — the prompt history sent to the provider's API — but it also includes the integrity of the code being written, the potential introduction of vulnerabilities suggested by the assistant, the intellectual property implications of training data used by the model, and the compliance implications if the assistant processes proprietary algorithms or regulated data without appropriate contractual protections. When that coding assistant is replaced by an AI agent with the ability to commit code, open pull requests, execute tests, and interact with production deployment pipelines, the attack surface expands to include the integrity of the entire software supply chain.

The pattern recurs across documented enterprise incidents. When an employee connects an unsanctioned agentic tool to enterprise systems using their own credentials — OAuth tokens, API keys, service account delegations — the resulting integration can be difficult to distinguish from authorized connections when viewed through standard network and identity monitoring, particularly where behavioral baselines for AI activity have not been established. No registration event triggers an inventory update; no approval process creates a governance record. The CSA and Token Security survey found that 82% of organizations discovered unknown AI agents in their environments in the past year, and that unauthorized integrations were among the most common vectors in AI agent security incidents [2]. The breach does not require a zero-day exploit; it exploits the gap between what the security program was designed to monitor and what an employee has actually deployed.

Multi-Cloud, Multi-Vendor Fragmentation

Enterprise AI deployments are not monolithic. The average enterprise uses AI capabilities from multiple foundation model providers, accessed through multiple cloud platforms, integrated through multiple orchestration frameworks, and consumed by multiple application teams with different governance maturity levels. The shared responsibility model that cloud providers established for infrastructure security — clearly delineating which security responsibilities belong to the provider and which belong to the customer — has not yet been coherently extended to AI systems in most enterprise environments.

The CSA AI Controls Matrix defines five supply chain roles in its Shared Security Responsibility Model: Cloud Service Provider (CSP), Model Provider (MP), Orchestrated Service Provider (OSP), Application Provider (AP), and AI Customer (AIC) [11]. Each role carries distinct control ownership obligations across AICM's 18 security domains. In a well-governed AI deployment, these responsibilities are explicitly assigned and evidence-based. In reality, most enterprises operate without clarity about where one role's responsibility ends and another's begins. When an enterprise deploys an AI application built on a foundation model from one provider, orchestrated through a framework from a second provider, hosted on infrastructure from a third provider, and accessed by employees through a SaaS interface from a fourth provider, the question of who is responsible for what security control across the stack becomes genuinely complex — and that complexity creates accountability voids that adversaries can exploit.

The Drift/Salesforce supply chain incident of August 2025 illustrates how multi-vendor fragmentation becomes an attack vector. Threat actor UNC6395 used stolen OAuth tokens from Drift's Salesforce integration to access customer environments across more than 700 organizations [10]. The attack succeeded in part because the security boundary between Drift's integration and customer Salesforce environments was not adequately governed — both by Drift, which had responsibility for securing the OAuth implementation, and by affected organizations, many of which had not evaluated the integration's access scope or monitored it for anomalous behavior. In a multi-vendor AI environment, every integration point where responsibility boundaries are unclear is a potential attack path.

Orphaned Agents: The Lifecycle Liability

Agent lifecycle governance represents perhaps the most underappreciated dimension of the ownership fragmentation problem. An AI agent that is deployed without formal governance is an agent that has no defined path to retirement. When the employee who deployed it leaves the organization, changes roles, or simply stops maintaining it, the organization is left with an orphaned agent: a persistent, potentially privileged autonomous system with no designated owner, no review cycle, no security update process, and no defined decommissioning trigger.

Strata's 2026 AI Agent Identity Crisis research found that only 23% of organizations have a formal, enterprise-wide strategy for agent identity management, while 37% rely on informal practices [12]. Only 21% maintain a real-time inventory of active agents, and nearly 80% of organizations deploying autonomous AI acknowledge that they cannot tell, in real time, what those systems are doing or who is responsible for them [12]. When organizations cannot answer basic inventory questions — how many agents are active, what credentials do they hold, what systems can they access, who owns them — the governance program cannot function regardless of how sophisticated the policy framework is.

The lifecycle risk compounds over time in a distinctive way. An agent that was never formally onboarded is unlikely to be formally retired. An agent whose access was provisioned as part of an initial deployment, rather than through a formal identity governance process, is unlikely to have its access scoped, reviewed, or

revoked on any predictable schedule. Over months or years, orphaned agents accumulate in enterprise environments, holding credentials to systems whose access was appropriate at deployment but may no longer be appropriate in the current organizational context. In the event of a security investigation, these agents create forensic complexity: when an autonomous workflow behaves unexpectedly, investigators need to know what triggered it, what systems it touched, and who owns it — precisely the information that lifecycle governance failures make difficult or impossible to reconstruct.

The Visibility Paradox

One of the more counterintuitive findings in recent enterprise AI security research appears in what the CSA and Token Security's 2026 survey report "Autonomous but Not Controlled" characterizes as the visibility paradox [2]. The majority of organizations — 68% — report that they have high visibility into their deployed AI agents and autonomous workflows. Yet in the same survey, 82% of organizations acknowledge that they discovered at least one AI agent or workflow in the past year that security or IT had not previously known about. These figures are not contradictory in the technical sense. They reveal something important about how security teams understand their own coverage: organizations believe they have visibility because they have deployed monitoring tools and established inventorying processes; they later discover they have visibility gaps because the tools and processes cover only what was formally deployed, not what was deployed outside the governance process.

This visibility gap is not primarily a tooling failure. It is a governance architecture failure. When AI agents and integrations can be provisioned by individual employees using corporate API keys, corporate OAuth accounts, or corporate SaaS credentials — without going through any process that would register them in an enterprise inventory — the monitoring layer is necessarily incomplete. Discovery after the fact, through periodic audits or security incidents, is not a governance control. It is an acknowledgment that governance has already failed.

Quantifying the Risk: From Hidden Assets to Realized Costs

The financial and operational consequences of shadow AI governance failures are now sufficiently documented to move beyond speculative risk assessment into data-grounded cost analysis. IBM's 2025 Cost of a Data Breach Report provides the most comprehensive shadow AI breach-cost data currently available in the public literature. Organizations IBM classifies as high shadow AI users incur breach costs approximately $670,000 more than those with minimal unauthorized AI deployment — $4.63 million versus $3.96 million, a 17% premium [1]. For context, the global average breach cost across all organizations stands

at $4.44 million [1]. In these breaches, personally identifiable information was compromised at a higher rate (65%) than in non-shadow-AI breaches (53%), and intellectual property appeared in 40% of shadow AI-related breaches [13]. Detection timelines are also longer: shadow AI breaches average 247 days to identify, compared to 241 days for standard breaches — a gap that reflects the inherent difficulty of detecting unauthorized systems that use legitimate credentials and behave in ways that can be difficult to distinguish from authorized integrations, particularly when behavioral baselines for the AI environment have not been established [13].

The cost of a shadow AI-linked breach is not only the direct data loss. It includes the regulatory exposure that follows. Organizations that have deployed AI tools without appropriate contractual protections, data processing agreements, or geographic data residency controls face potential regulatory action that is independent of whether a breach occurred. In healthcare, 57% of professionals have encountered or used unauthorized AI tools, often processing protected health information without Business Associate Agreements, creating HIPAA liability exposure of up to $1.5 million per violation category [14]. Industry analyses interpret established HIPAA enforcement principles as extending accountability to covered entities and business associates regardless of whether an AI deployment was formally authorized — meaning that the unauthorized nature of a deployment does not constitute a compliance defense [14]. The EU AI Act's high-risk AI obligations take effect in August 2026, creating a new layer of regulatory exposure for organizations that cannot demonstrate governance over their AI deployments [22]. Colorado's AI Act becomes enforceable in June 2026 [15].

Beyond direct financial costs, the operational impact of shadow AI governance failures includes the disruption cost of AI agent incidents. The CSA and Token Security survey found that of the 65% of organizations that experienced an AI agent security incident in the past year, the most commonly reported impacts were data exposure (61%) and operational disruption (43%) [2]. Operational disruption from an ungoverned agent — one that takes unexpected actions in production systems, corrupts data pipelines, or generates outputs that propagate through downstream workflows before detection — can be difficult to contain and expensive to remediate, particularly when the agent's action history cannot be reconstructed.

Attack Surface Anatomy: How Fragmentation Exploits Each Layer

The MAESTRO framework, developed by CSA for agentic AI threat modeling, provides a useful lens for understanding how ownership fragmentation creates attack surface at each layer of the AI stack [16]. MAESTRO's seven-layer architecture — Foundation Models, Data Operations, Agent Frameworks, Deployment and Infrastructure, Evaluation and Observability, Security and Compliance, and Agent Ecosystem — makes visible how a governance gap at one layer enables exploitation at another.

At the Foundation Model layer, shadow AI deployments typically mean that the organization has not evaluated the model provider's security posture, data handling practices, or training data provenance. Employees submitting sensitive organizational data to an unapproved model provider have no assurance that the provider does not use submission data for model retraining, that the provider's infrastructure meets the organization's security standards, or that the provider's contractual obligations satisfy the organization's regulatory requirements. The attack surface here is not primarily adversarial in the traditional sense — it is structural: data exits the organization's control into an environment whose security properties have not been assessed.

At the Data Operations layer, shadow AI creates risks around what data is being processed by AI systems and what is being generated as a result. Shadow AI integrations commonly process production data — customer records, employee information, financial data, intellectual property — without the data governance controls that formal deployment would require. Output data from shadow AI systems may be incorporated into business processes, stored in systems with different retention policies, or shared with additional services without any audit trail that connects the output to its source.

At the Agent Frameworks layer, the attack surface expands dramatically. Employees deploying agent frameworks such as LangChain or CrewAI are, in most cases, granting those frameworks access to corporate systems through service accounts, API keys, or OAuth connections provisioned through their own enterprise credentials. The security scope of what these frameworks can access often extends well beyond what the deploying employee has explicitly authorized — because agent frameworks, by design, discover and utilize available tools and connections. An agent deployed with access to a corporate email account may, through that access, be able to read the employee's calendar, contact list, and all historical correspondence, including communications that contain sensitive organizational information never intended to be processed by an AI system.

At the Deployment and Infrastructure layer, shadow agents deployed on personal laptops, personal cloud accounts, or the employee's development machine create infrastructure security gaps that endpoint and network security tools may not capture. A CrewAI workflow running on an employee's personal cloud account, accessing corporate systems through legitimate credentials, generates network traffic that appears authorized and endpoints that appear compliant while the underlying governance controls — logging, monitoring, access review, incident response — are entirely absent.

At the Evaluation and Observability layer, the visibility paradox is most acute. When an agent is not formally inventoried, the observability tooling that watches the enterprise environment has no baseline for what that agent's normal behavior looks like. It cannot distinguish anomalous agent behavior from normal agent behavior because it has no record of either. It cannot initiate an investigation when the agent's behavior changes because it has no alert configured for an asset it does not know exists. It cannot reconstruct what the agent did during a security incident because the agent's action history may not have been logged, or may have been logged in the agent's own operational store rather than in the enterprise's security information and event management infrastructure.

The Security and Compliance layer and Agent Ecosystem layer complete the picture. Shadow AI systems that operate without security review have no formal place in the organization's incident response plan. When they are implicated in a security event, the response team must simultaneously investigate the incident and reverse-engineer the governance history of the asset — who deployed it, when, with what access, and through what chain of credentials. In complex agentic environments, where agents may communicate with other agents and spawn sub-agents, reconstructing this history from scratch during an active incident is both difficult and time-consuming.

Regulatory Exposure: Compliance in the Dark

The regulatory environment for enterprise AI deployments is maturing faster than many enterprises' governance programs. Organizations that have tolerated shadow AI deployments as a managed ambiguity are increasingly exposed to regulatory frameworks that presuppose the ability to demonstrate governance over all AI systems, not only the ones that went through formal review.

The EU AI Act's tiered risk classification requires that organizations deploying high-risk AI systems — defined by application domain rather than by the organization's internal classification of the deployment — maintain technical documentation, automatic logging, human oversight mechanisms, and quality management systems [22]. An organization that has deployed an AI system in a high-risk application domain without going through the formal assessment process cannot produce the technical documentation the Act requires, because the documentation was never created. The compliance gap is not addressable retrospectively in the way that some compliance programs permit; it requires either re-establishing the governance record for the system or removing it from service. Neither option is straightforward for a shadow AI deployment that has become embedded in operational workflows.

Colorado's Artificial Intelligence Act, effective June 2026, establishes requirements for high-risk AI systems affecting Colorado consumers, including impact assessments and governance documentation [15]. HIPAA enforcement in the United States increasingly scrutinizes AI processing of protected health information, particularly in light of established OCR principles that extend covered entity responsibility to all processing activities — authorized or otherwise — conducted within the organization's operational scope [14].

These regulatory obligations create a practical problem for organizations with shadow AI environments: compliance audits are beginning to ask not only about authorized AI deployments but about controls designed to discover and remediate unauthorized ones. An organization that cannot demonstrate a shadow AI discovery and governance program will increasingly face compliance findings that reflect not only specific violations but a systemic governance failure affecting regulatory confidence in the organization's entire AI posture.

A Framework for Closing the Governance Gap

The governance response to ownership fragmentation must be structural rather than symptomatic. Issuing a policy document that prohibits shadow AI and providing a list of approved tools addresses the compliance record without addressing the conditions that produce shadow AI in the first place. A structural response requires four elements: clear ownership assignment at every layer, a shadow AI discovery program capable of continuous inventory, agent lifecycle governance that spans provisioning to decommissioning, and a control substrate that makes the governance program auditable and evidence-based.

Establishing Clear Ownership

Every AI system in the enterprise environment — whether formally procured, organically adopted, or actively migrated from shadow to governed status — requires a designated owner who has both accountability for the system's security posture and the authority to direct remediation when that posture degrades. The challenge is that ownership in the AI supply chain is, in practice, typically distributed across multiple parties. The CSA AICM's Shared Security Responsibility Model provides a practical framework for assigning this ownership [11]. By mapping each control obligation across the five SSRM roles — CSP, MP, OSP, AP, and AIC — organizations can identify exactly which governance responsibilities they bear as AI customers, which they share with the application provider, and which they can reasonably expect the model provider to own. This mapping does not eliminate the complexity of multi-vendor AI governance; it makes that complexity legible and manageable.

At the organizational level, ownership clarity requires an explicit governance model that specifies which executive function is accountable for AI security risk, which function has authority over AI procurement and deployment approvals, and which function is responsible for ongoing monitoring and incident response. This paper recommends that the CISO function be accountable for AI security risk management, with formal integration into procurement and deployment processes — not merely notification after the fact. Without authority to shape the processes through which AI enters the enterprise environment, CISO accountability for AI security risk is nominal rather than operational.

Building a Shadow AI Discovery Program

Discovery is not a one-time exercise. It is an ongoing operational capability. The scale and velocity of shadow AI deployment — a 509% increase in shadow AI tool usage was reported in one 2026 analysis [18] — exceeds what periodic manual audits can track. An effective shadow AI discovery program combines several complementary capabilities: network traffic analysis to identify data flows to known AI provider endpoints; identity and access management review to identify service accounts and API keys provisioned for AI integrations; SaaS discovery tooling extended to cover AI-enabled applications; and agent behavior

monitoring to identify automated workflows that exhibit agent-like characteristics (sequential API calls, decision branching, autonomous document modification) without a corresponding entry in the formal agent inventory.

When shadow AI assets are discovered, the governance response requires a defined triage process. Not every shadow AI deployment represents equal risk. An employee using an approved AI model's web interface through a personal account to draft documents is a different risk profile than an AI agent with write access to the company's data warehouse that has been running unmonitored for six months. The triage process should assess the sensitivity of the data the system has accessed, the scope of the actions it is capable of taking, the presence or absence of a legitimate business purpose, and the feasibility of migrating it to a governed deployment. The outcome of triage should be either remediation — migration to governed status, access restriction, or decommissioning — or documented acceptance of residual risk by an accountable owner.

Agent Lifecycle Governance

Agent lifecycle governance begins at provisioning, not at discovery. An enterprise that requires formal registration of every AI agent deployment, as a condition of granting the API credentials or service account access the agent needs to function, builds discovery into the deployment process rather than relying on detection after the fact. Microsoft's Agent Governance Toolkit, released in April 2026, represents one emerging model for agent lifecycle governance tooling that integrates identity assignment, permission scoping, and lifecycle tracking into a unified management plane [19]. The OWASP Top 10 for Agentic Applications 2026, published in December 2025, identifies unauthorized agent actions, agent identity compromise, and agent lifecycle failures as among the most significant agentic risks [20].

A practical lifecycle governance program includes four phases. Provisioning establishes the agent's formal identity, assigns an organizational owner, scopes the agent's access to the minimum permissions required for its stated function, and registers the agent in the enterprise's AI inventory. Operation subjects the agent to continuous monitoring against a behavioral baseline, with alerting configured for anomalous actions. Review applies a periodic recertification process — analogous to access certification in identity governance — that requires the designated owner to affirm that the agent's continued operation, access scope, and behavioral profile remain appropriate. Decommissioning revokes the agent's credentials, removes its integrations, archives its action history, and updates the enterprise inventory to reflect the retirement. Every phase requires a designated accountable owner; without clear ownership, lifecycle governance processes fail at the transitions.

AICM as the Control Substrate

The governance framework described above requires an underlying control architecture that makes ownership assignments, discovery requirements, and lifecycle obligations operationally concrete and auditable. The CSA AI Controls Matrix v1.0.3 provides that architecture [11]. Across its 18 domains and 243 control objectives, AICM maps the full operational range that enterprise AI governance requires, from data security and identity management to supply chain transparency and logging. For shadow AI governance specifically, several AICM domains carry particular weight.

The Governance, Risk Management and Compliance (GRC) domain's 15 controls establish the organizational governance structures — AI governance programs, risk registers, policy frameworks, regulatory alignment mechanisms — that make accountability assignments durable rather than nominal. The Identity and Access Management (IAM) domain's 19 controls address access controls for AI models, inference APIs, and training pipelines, providing the control framework within which agent identity and access governance operates. The Logging and Monitoring (LOG) domain's 15 controls establish the AI behavior telemetry, audit logging, and anomaly detection capabilities that make the visibility paradox tractable — not by eliminating the gap between stated and actual visibility, but by instrumenting the environment so that shadow AI discovery becomes a continuous operational output rather than a periodic investigative exercise. The Supply Chain Management Transparency and Accountability (STA) domain's 16 controls address third-party model risk, AI vendor assessment, and AI bill of materials, providing the framework for multi-vendor accountability.

The AI-CAIQ, the structured self-assessment questionnaire that maps directly to AICM controls, provides a practical starting point for organizations that need to rapidly assess their current state against these control objectives. By completing the AI-CAIQ for the AI systems within their governance scope — and, critically, by extending that scope to include shadow AI assets discovered through the discovery program — organizations can produce an evidence-based picture of control gaps that is both internally actionable and externally auditable through the CSA STAR for AI registry.

CSA Resource Alignment

The framework described in this paper maps directly to several CSA resources that provide more detailed guidance on specific dimensions of shadow AI governance.

The **CSA AI Controls Matrix (AICM) v1.0.3** provides the control catalog that underlies the governance framework described here. AICM's 243 controls across 18 domains, organized by supply chain role and lifecycle phase, give organizations the operational specificity that higher-level governance frameworks lack. The SSRM within AICM is particularly relevant for addressing multi-vendor accountability fragmentation [11].

The **MAESTRO Agentic AI Threat Modeling Framework** provides the threat analysis methodology that this paper applies to shadow AI environments. MAESTRO's seven-layer architecture makes AI system threat modeling systematic and reproducible, enabling organizations to evaluate shadow AI deployments against a common threat taxonomy before or in parallel with migration to governed status [16][17].

The **CSA AI Organizational Responsibilities: Governance, Risk Management, Compliance and Cultural Aspects** publication directly addresses shadow AI prevention as a governance design problem. Its treatment of how to structure organizational accountability for AI, define clear decision rights, and embed AI governance into existing GRC programs provides practical organizational guidance that complements the technical control framework of AICM [21].

The **"Autonomous but Not Controlled" CSA and Token Security Survey Report** provides empirical data on the state of enterprise AI agent governance, including the visibility paradox data cited in this paper. Its findings on agent discovery rates, incident prevalence, and governance maturity serve as a baseline against which organizations can assess their own programs [2].

The **CSA STAR for AI Registry** enables organizations to publish AI-CAIQ responses, making their AI governance posture visible to customers, partners, and regulators. For organizations seeking to differentiate their governance maturity in an environment where shadow AI concerns are now a routine element of enterprise vendor risk assessments, STAR for AI participation provides a credible, evidence-based disclosure mechanism.

The **CSA AI Safety Initiative** provides the broader research and standards context within which the shadow AI governance problem should be understood. The convergence of shadow AI with agentic AI, with multi-agent architectures, and with the increasing integration of AI systems into critical business functions represents a safety challenge, not only a security challenge. Governance frameworks designed only for the breach risk dimension of shadow AI will be insufficient for the operational safety dimension as AI systems acquire greater autonomy and broader action scope.

Conclusions and Recommendations

Shadow AI is not a fringe phenomenon. It is the normalized state of AI deployment in most enterprises, coexisting alongside formal governance programs that cover only a fraction of the AI systems actually in use. The data surveyed in this paper consistently indicates that the majority of organizations operate with the majority of their AI ecosystem outside formal governance scope, that this gap is a direct contributor to security incidents and breach costs, and that the gap is widening as agentic AI deployments accelerate faster than governance program development.

The policy response — prohibiting unauthorized AI, providing approved tool lists, training employees — is necessary but not sufficient. It addresses the symptom (individual behavior) without addressing the cause (governance architecture). Organizations that want to close the shadow AI governance gap must address ownership fragmentation structurally, which requires the following actions.

Establish an accountable AI governance function with real authority over AI procurement, deployment, and lifecycle management. This function should own the enterprise's AI inventory, operate the shadow AI discovery program, and have the authority — not merely the responsibility — to govern what enters the AI environment. Without authority proportionate to accountability, governance programs produce documentation without capability.

Deploy continuous shadow AI discovery as an operational function, not an annual audit exercise. The velocity of AI adoption demands near-real-time inventory capabilities. Network traffic analysis, identity governance review, and SaaS discovery should be coordinated to produce a continuously updated picture of what AI systems are operating in the enterprise environment and who owns them.

Implement agent lifecycle governance before the next generation of agentic deployments makes the problem intractable. The governance cost of establishing lifecycle management processes now — provisioning registration, access scoping, behavioral monitoring, periodic recertification, defined decommissioning — is substantially lower than the remediation cost of governing a sprawling ecosystem of orphaned agents after a security incident exposes their existence and their access scope.

Apply the AICM control framework to shadow AI governance as well as formally procured AI systems. AICM's control architecture was designed for the full range of enterprise AI deployments, not only those that went through security review. Using AICM's GRC, IAM, LOG, and STA domains as the baseline for shadow AI governance conversations gives security teams a common vocabulary and a clear accountability structure for the difficult cross-functional conversations that effective AI governance requires.

Engage with emerging regulatory requirements proactively. The EU AI Act, Colorado AI Act, and HIPAA AI guidance are the leading edge of a regulatory tide that will raise the compliance cost of governance failures over time. Organizations that establish shadow AI discovery and governance programs now will be better positioned to demonstrate compliance posture when regulatory scrutiny increases — and to avoid the most serious consequences of governance failures that are increasingly difficult to characterize as unforeseeable.

Shadow AI is not going away. The economic incentive for employees to use the most capable available tools, the organizational incentive for business units to adopt AI capabilities that improve their competitive position, and the technical accessibility of powerful AI systems through consumer interfaces are all structural features of the current environment that governance programs cannot eliminate. The appropriate goal is not zero shadow AI. It is a governance architecture capable of discovering what exists, assessing what

it risks, and bringing high-risk deployments into managed status before they become security events. The organizations that build that architecture now will be better prepared for the AI security landscape that the evidence strongly suggests is already arriving.

References

[1] IBM Security. "Cost of a Data Breach Report 2025." IBM, July 2025.

[2] Cloud Security Alliance and Token Security. "Autonomous but Not Controlled: AI Agent Incidents Now Common in Enterprises." CSA, 2026.

[3] JumpCloud. "11 Stats About Shadow AI in 2026." JumpCloud, 2026.

[4] Help Net Security. "Shadow AI risks deepen as 31% of users get no employer training." Help Net Security, May 2026.

[5] CIO. "Roughly half of employees are using unsanctioned AI tools, and enterprise leaders are major culprits." IDG Communications, 2026.

[6] Acuvity. "2025 State of AI Security Report: Application Risk Findings." BusinessWire, October 2025.

[7] Acuvity. "2025 State of AI Security Report: Governance Ownership Findings." BusinessWire, October 2025.

[8] Fortium Partners. "Beyond the CAIO: Defining Executive Accountability for AI Risk in the Modern C-Suite." Fortium Partners, 2026.

[9] Cybersecurity Insiders. "2026 CISO AI Risk Report on Cybersecurity Risks." Cybersecurity Insiders, 2026.

[10] Reco AI. "AI & Cloud Security Breaches: 2025 Year in Review." Reco AI, 2025.

[11] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.0.3." CSA, 2025.

[12] Strata Identity. "The AI Agent Identity Crisis: A 2026 Guide." Strata Identity, 2026.

[13] Kiteworks. "How Shadow AI Costs Companies $670K Extra: IBM's 2025 Breach Report." Kiteworks, 2025.

[14] Netwrix. "12 Critical Shadow AI Security Risks Your Organization Needs to Monitor in 2026." Netwrix, 2026.

[15] Giovanni Coletta. "Agentic AI Governance Frameworks 2026: Emerging Standards, Risks and Insights." Medium, 2026.

[16] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA, February 2025.

[17] Cloud Security Alliance. "MAESTRO for Real-World Agentic AI Threats." CSA, February 2026.

[18] PR Newswire. "Cyberhaven Solves the 509% Surge in Shadow AI." Cyberhaven, 2026.

[19] Microsoft Open Source Blog. "Introducing the Agent Governance Toolkit: Open-source runtime security for AI agents." Microsoft, April 2026.

[20] OWASP Gen AI Security Project. "OWASP Top 10 for Agentic Applications for 2026." OWASP, December 2025.

[21] Cloud Security Alliance. "AI Organizational Responsibilities: Governance, Risk Management, Compliance and Cultural Aspects." CSA, 2024.

[22] European Parliament and Council. "Regulation (EU) 2024/1689 on Artificial Intelligence (EU AI Act)." Official Journal of the European Union, July 2024.

