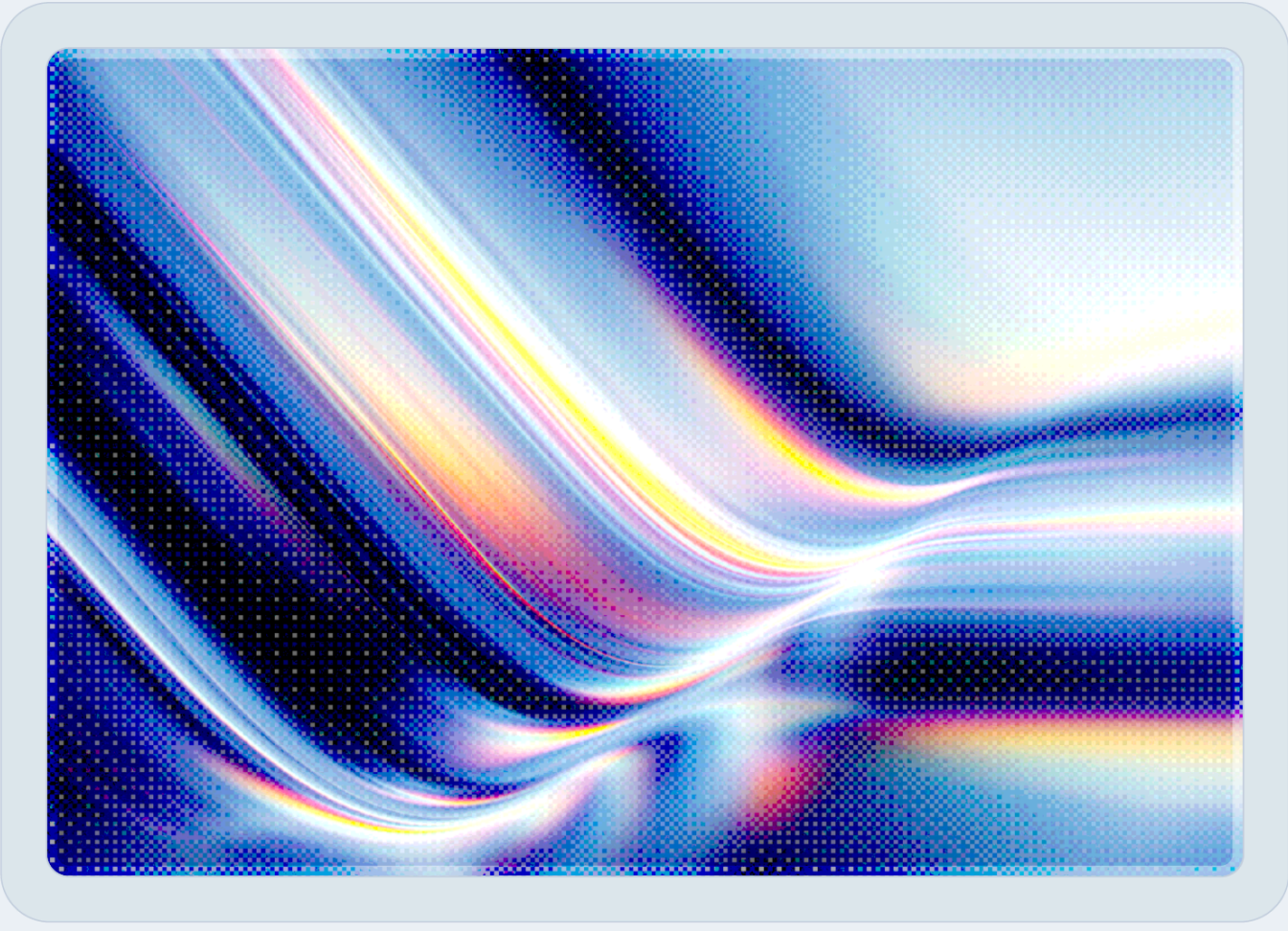
CSAI Foundation | Cloud Security Alliance

# Shadow AI and the Enterprise Visibility Crisis

## Confronting Structural Blindness to the AI Attack Surface

2026-05-28

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

# Table of Contents

# Executive Summary

Enterprise AI adoption has substantially outpaced enterprise AI governance, creating a gap that security leaders consistently rank among their top concerns. Ninety-eight percent of organizations now report unsanctioned AI use within their workforce [19], yet 86% lack visibility into how data flows to and from AI tools [2], and only 37% have any AI governance policy in place [1]. The gap between these numbers is not a policy problem or a culture problem. It is a structural problem: the security architectures, discovery tools, and governance processes that organizations built for traditional software and cloud services were not designed to detect, classify, or control AI systems.

The consequences of this structural blindness are already visible in loss data. Shadow AI has been identified as a contributing factor in one in five data breaches [19], adding an average of $670,000 to incident costs [17]. The average enterprise now experiences 223 data policy violations per month related to AI usage [17]. A 2025 survey conducted by the Cloud Security Alliance and Google Cloud found that 72% of respondents lacked confidence in their organization's ability to execute an AI security strategy [20], even as AI adoption accelerated across the same organizations.

The shadow AI problem has three distinct but interrelated dimensions. The first is the familiar story of unsanctioned standalone tools—employees using ChatGPT, Gemini, or other consumer-facing AI products to process work data outside enterprise controls. The second is subtler and more dangerous: AI capabilities embedded within already-approved SaaS applications. When an enterprise licenses Salesforce, Slack, Microsoft 365, or GitHub, the security review that approved those platforms often predates the AI features that have since been added to them. Employees may be interacting with production AI capabilities that have never been through an enterprise risk assessment. The third dimension is the most urgent: the proliferation of autonomous AI agents—systems that act, not just answer—deployed with minimal oversight across enterprise environments. Gravitee's State of AI Agent Security 2026 report estimates that more than 3 million AI agents are now operating within corporations, with fewer than half actively monitored or secured [4].

This whitepaper argues that visibility is the prerequisite for every other AI security control. Organizations cannot enforce policy on systems they cannot see, cannot assess risk in environments they have not inventoried, and cannot detect anomalous behavior from agents whose baseline behavior was never established. The path forward begins with treating AI assets with the same rigor that security teams apply to cloud workloads: systematic discovery, risk classification, continuous monitoring, and lifecycle governance. The Cloud Security Alliance has developed a suite of frameworks—the AI Controls Matrix, the MAESTRO agentic AI threat modeling framework, the Shadow Access research series, and the AI Organizational Responsibilities guidance—that together provide the conceptual and operational foundation for this work.

# 1. Introduction: The Governance Architecture Has Not Caught Up

The history of enterprise security is, in part, a history of governance catching up to adoption. When employees began uploading files to Dropbox and Google Drive ahead of IT-sanctioned alternatives, the industry developed cloud access security broker (CASB) technology and shadow IT discovery capabilities. When containerization outpaced infrastructure governance, security teams built runtime protection and image scanning into CI/CD pipelines. In each case, the pattern was the same: adoption created a visibility gap, the visibility gap became a security gap, and eventually tooling and process closed the gap—usually after some number of painful incidents.

AI adoption is replicating this pattern at a speed and breadth that have outpaced most prior enterprise technology transitions, including the initial waves of cloud and mobile adoption. According to Lenovo research cited in industry surveys, 70% of enterprise AI use occurs outside IT oversight [2]. Nearly half of generative AI users access tools through personal accounts, entirely bypassing enterprise controls and monitoring [5]. When employees do use enterprise-approved platforms, they may be interacting with AI features that were never reviewed as part of the original procurement decision, because those features did not exist when the contract was signed.

The enterprise security function is not ignorant of this situation. Security leaders consistently rank shadow AI as a top concern, and the CSA/Google Cloud State of AI Security and Governance survey found that 70% of respondents report at least moderate executive awareness of AI security implications [20]. The gap is not awareness. The gap is the absence of discovery mechanisms, inventory frameworks, and monitoring architectures capable of surfacing AI assets across the full scope of enterprise environments.

Shadow AI is not simply a new category of shadow IT. Traditional shadow IT—a rogue SaaS subscription, an employee-installed application—occupied a defined position in the enterprise architecture: a software asset with a data footprint that CASB and endpoint tools could, in principle, detect. AI systems are different in kind. They sit inside applications, operate through APIs that blend with legitimate traffic, execute autonomously via agent frameworks, and leave audit trails that existing SIEM and DLP tooling was not designed to parse. Understanding why enterprises are structurally blind to AI requires understanding the three forms that shadow AI actually takes.

# 2. The Three Dimensions of Shadow AI

## 2.1 Unsanctioned Standalone Tools

The most visible form of shadow AI—and the one that has received the most policy attention—is the direct employee use of consumer-facing AI products without enterprise approval or oversight. Employees use ChatGPT to draft communications, GitHub Copilot to generate code, Gemini to summarize documents, and an expanding catalog of vertical AI tools for tasks ranging from contract review to market analysis. The data they feed into these interactions often include proprietary business information, customer records, source code, and financial data.

LayerX research, as compiled in industry surveys, found that 77% of enterprise AI users regularly copy and paste data into AI chatbots, and that 22% of those interactions contain personally identifiable information or payment card data [2]. Forty percent of file uploads to generative AI platforms include PII or PCI-relevant content [2]. These are not malicious acts; they are the ordinary behavior of knowledge workers applying available tools to their workloads. The security exposure is structural, not intentional—and no amount of policy communication resolves a structural problem.

The security implication is not confined to data exfiltration risk. When an employee processes sensitive data in a consumer AI service, the enterprise loses chain-of-custody control. Training data contamination, inference attacks, and third-party data retention all become relevant threat vectors, with exposure that may not materialize as an incident for months or years. The absence of an audit trail compounds the problem: security teams responding to a breach involving AI-processed data often cannot reconstruct what was processed, when, or in what context.

## 2.2 AI Embedded in Sanctioned SaaS Applications

The second dimension of shadow AI is less visible but arguably more consequential, because it operates entirely inside the perimeter of enterprise-approved systems. Virtually every major SaaS platform has added generative AI capabilities since 2023. Microsoft 365 Copilot operates across Exchange, SharePoint, Teams, and OneDrive, with access to the full scope of whatever data the user can access. Salesforce Einstein synthesizes CRM data. Slack AI summarizes conversation threads. Notion AI assists with document drafting. In each case, the AI feature may have been activated without a security review, either because it was included in an existing license upgrade or because an administrator enabled it without triggering a formal risk assessment.

This creates what might be called the embedded AI problem: AI capabilities with access to the full data footprint of enterprise-approved platforms, operating without the oversight applied to the platforms themselves. The security review that approved Microsoft 365 in 2022 addressed Microsoft's cloud security architecture, access controls, and data handling terms for email and collaboration. It did not address a generative AI layer that did not yet exist. Extending an existing approval to cover fundamentally new AI capabilities requires deliberate reassessment—a step that many organizations have not taken.

The embedded AI problem is compounded by AI features added to developer tools. A high-severity vulnerability in GitHub Copilot (CVE-2025-59145, CVSS 9.6), disclosed in late 2025, demonstrated that attackers could embed hidden instructions in pull request descriptions that Copilot would execute on behalf of code reviewers, silently exfiltrating source code, API keys, and secrets from private repositories [13]. The attack technique, named CamoLeak, exploited the embedded context-access that made Copilot useful—and turned it into an exfiltration channel. This class of attack, prompt injection via embedded AI, is a structural risk in any AI system that processes untrusted input with elevated data access.

## 2.3 Autonomous Agent Proliferation

The third dimension of shadow AI is the fastest-growing and the least governed: autonomous AI agents. Unlike AI assistants, which respond to queries, agents act: they plan sequences of steps, invoke tools and APIs, write and execute code, manage files, and interact with external systems—all with minimal or no human oversight at the task level. The enterprise AI ecosystem is generating agents faster than governance frameworks can accommodate them [14].

Gartner predicts that 40% of enterprise applications will feature task-specific AI agents by the end of 2026, up from fewer than 5% in 2025 [3]. Gartner further projects that by 2028, the average Fortune 500 enterprise will have more than 150,000 agents in use, up from fewer than 15 in 2025 [3]. Gravitee's 2026 research found that more than 3 million AI agents are already operating within corporations, with only 47.1% actively monitored or secured—leaving approximately 1.5 million agents running without oversight, accessing sensitive data, making consequential decisions, and connecting to critical systems with no audit trail [4].

The governance gap these statistics reveal is significant. Gravitee found that 81% of enterprise AI teams are past the planning phase for agent deployment, yet only 14.4% have received full security approval for their agent implementations [4]. Only 22% of teams treat agents as independent identities—a prerequisite for any meaningful access control—with most organizations relying instead on shared API keys that provide no granular auditability [4]. This creates conditions in which agents accumulate access rights, interact with sensitive systems, and generate outputs that feed downstream processes, entirely outside the visibility of security teams.

# 3. Why Enterprises Are Structurally Blind

## 3.1 Discovery Tooling Built for a Different Problem

The core of the enterprise AI visibility problem is architectural: existing security tooling was designed to detect software assets in defined locations, and AI is not that kind of asset. Traditional shadow IT discovery works by identifying network connections to unauthorized domains, cataloging installed applications on managed endpoints, or parsing SaaS usage logs for unrecognized service names. These approaches find AI use in some cases—a spike in traffic to api.openai.com, an installed desktop application—but they are systematically blind to the majority of AI use patterns.

AI embedded in SaaS platforms does not generate distinct network traffic that differs from legitimate SaaS usage. An employee querying Salesforce Einstein generates API calls to Salesforce endpoints, which appear in logs identically to any other Salesforce interaction. AI agents built on cloud platform services—AWS Bedrock, Azure OpenAI, Google Vertex AI—may be indistinguishable at the network level from other enterprise cloud workloads. Browser-based AI tools accessed through a sanctioned browser leave no endpoint artifact at all. Gartner has characterized existing DLP and CASB tools as having critical generative AI blind spots, as they were not architected for prompt-level inspection or embedded AI tool discovery [3].

The implication is that organizations applying traditional shadow IT discovery to the AI problem will systematically undercount AI exposure. They will find the obvious cases—direct traffic to consumer AI services—and miss the majority of risk: embedded AI features, API-based agent interactions, and browser-based tool use that generates no distinguishable signal.

## 3.2 Non-Human Identity Invisibility

AI agents interact with enterprise systems by assuming identities: service accounts, API keys, OAuth tokens, and delegated credentials that grant them access to the systems and data they need to function. Non-human identities have historically been a governance blind spot for most enterprises—the Cloud Security Alliance's research on shadow access has documented the widespread proliferation of over-permissioned, under-monitored non-human identities as a foundational cloud security challenge [7]. AI agents dramatically amplify this existing problem.

An agent that performs calendar management, drafts email, queries a CRM, and writes to a project management tool requires credentials across all of those systems. If those credentials are managed as shared API keys—the current practice in the majority of enterprises, according to Gravitee [4]—they are invisible to IAM governance processes that focus on human identity lifecycle management. There is no

joiner-mover-leaver process for API keys. There is no access recertification for service accounts created by a developer building an agent proof-of-concept. There is no MFA for an OAuth token granted to an agent that has been accumulating permissions over six months of iterative development.

The CSA's Shadow Access research series defines this category of risk as unmonitored, unauthorized, invisible, unsafe, and over-permissioned cloud access [7][8]. For AI agents, the definition fits with particular precision: agents that accumulate permissions over their lifecycle, that are never subjected to access review, and that continue to hold credentials long after the use case that justified them has changed or been abandoned, represent exactly the over-permissioned shadow access profile that has characterized major identity-based attacks.

## 3.3 The Approval Gap in SaaS AI Features

When an enterprise licenses a SaaS platform, the procurement and security review process establishes what the platform does and what data it will access. Those determinations are recorded in vendor risk assessments, data processing agreements, and security questionnaires. The challenge with AI features embedded in SaaS platforms is that they routinely change what the platform does and what data it accesses—after the review is complete.

Microsoft 365 Copilot, for example, operates on the principle of accessing everything a user can access, which in many organizations means significantly more than IT governance processes assumed. Organizations routinely discover oversharing problems only after enabling Copilot—because the AI's ability to surface and synthesize information across the entire SharePoint and OneDrive footprint reveals data access paths that human users rarely followed. The problem was not created by Copilot; it was revealed by Copilot. But the revelation comes after deployment, not before—precisely because existing security review processes were not designed to ask AI-specific questions.

The same approval gap applies to AI features in developer tools, data analytics platforms, customer support systems, and HR software. In each domain, an AI layer has been added to a platform whose data access permissions were set for human use patterns. Those permissions, when exercised by an AI system at machine speed and scale, create exposure profiles that human access governance frameworks did not anticipate.

# 4. The Threat Surface That Blindness Creates

## 4.1 Data Exfiltration and Unintended Disclosure

The most immediate consequence of shadow AI is data exfiltration—not in the sense of a sophisticated attack, but in the ordinary sense of sensitive data leaving enterprise control through channels that IT and security did not intend, monitor, or approve. Every time an employee pastes a client contract into a consumer AI chatbot, sends a proprietary product specification to a code generation service, or uploads a financial model to an AI summarization tool, enterprise data crosses a boundary that the data classification and handling policies never addressed, because those policies were written before such tools existed.

The volume of these ordinary exfiltration events is substantial. The average enterprise's 223 monthly data policy violations related to AI [17] are not the work of malicious insiders. They are the aggregate effect of a workforce applying available productivity tools to real work, in the absence of sanctioned alternatives that meet the same productivity need. Blocking without providing sanctioned alternatives has repeatedly failed as a standalone governance strategy—the productivity need that drives shadow AI use does not disappear when access is restricted. Research suggests that when enterprises provide approved AI alternatives, unauthorized use drops by as much as 89% [1]. The inverse—maintaining restrictions without alternatives—achieves some reduction in detectable violations while driving a larger share of AI use to channels that generate no detection signal at all.

Beyond ordinary disclosure, shadow AI creates conditions for targeted data extraction. An adversary who achieves prompt injection access to an AI system with broad data permissions—whether through a malicious document, a poisoned knowledge base, or an adversarial instruction embedded in content the agent processes—can exfiltrate data at a scale and specificity that manual access would require significant time and skill to replicate. The CamoLeak vulnerability in GitHub Copilot demonstrated this attack class: the AI system's broad access to repository content, combined with its tendency to follow instructions embedded in the content it processes, created an exfiltration channel that bypassed traditional DLP controls entirely [13].

## 4.2 AI-Mediated Attack Chains

Shadow AI creates attack surface not only through data exposure but through the autonomous action capabilities of AI agents. An agent that can read email, schedule meetings, write and commit code, query databases, and call external APIs is a powerful autonomous actor. An adversary who can influence that agent's behavior—through prompt injection, poisoned training data, or manipulation of the knowledge sources the agent consults—gains the ability to direct that autonomous action toward malicious ends.

This attack class falls within the threat categories addressed by the CSA's AI Controls Matrix [10] and is particularly dangerous in shadow AI contexts because the absence of monitoring means that anomalous agent behavior may go undetected for extended periods. An agent that has been manipulated to exfiltrate data over time, to modify code in ways that introduce vulnerabilities, or to forward sensitive communications to external addresses will produce no alert in a security environment that has no baseline for that agent's normal behavior, because no baseline was ever established.

Agentic AI also introduces a new class of privilege escalation risk. An agent operating with permissions appropriate to its intended function may, through a sequence of individually permitted API calls, achieve access to resources or capabilities that no single call would have reached. This chained access vulnerability, analyzed in the CSA's Shadow Access and AI research [9], represents a structural challenge for access governance frameworks designed around point-in-time permission evaluation rather than sequence-aware reasoning.

## 4.3 Regulatory and Compliance Exposure

The compliance dimension of shadow AI is rapidly becoming concrete. The European Union AI Act requires organizations to maintain a complete inventory of all deployed AI systems classified by risk tier; enforcement obligations for high-risk AI systems take effect on August 2, 2026 [18]. These obligations apply regardless of how a system was adopted—there is no exemption for tools that employees adopted without IT approval. An organization that cannot enumerate its AI systems cannot demonstrate compliance with a regulation that begins with enumeration as a prerequisite.

The EU AI Act is not alone. Across major regulatory jurisdictions, AI-specific governance obligations are accumulating: requirements for AI system documentation, impact assessments, human oversight mechanisms, and audit trails. Shadow AI is, by definition, the absence of the documentation, assessment, oversight, and audit trail that these requirements demand. The regulatory exposure is not hypothetical. Gartner predicts that AI-related legal claims will exceed 2,000 by the end of 2026 as a direct result of insufficient governance guardrails [3], and that by 2030, more than 40% of enterprises will face security or compliance incidents stemming directly from unauthorized AI use [3].

For organizations subject to sector-specific regulations—financial services, healthcare, government—the compliance exposure is amplified by existing data handling requirements that AI systems may violate in ways that neither the employee using the tool nor the security team monitoring the environment is positioned to detect.

# 5. The Measurement Problem

Enterprise security has long operated on the principle that you cannot secure what you cannot see. For AI, the more pressing reality is that organizations do not yet know the full scope of what they cannot see—and their estimates of shadow AI exposure are almost certainly undercounts.

The CSA/Google Cloud State of AI Security and Governance survey found that only 26% of organizations have comprehensive AI security governance policies, while 64% have some guidelines or are in the process of developing them [20]. Even organizations with formal governance policies may not have the discovery mechanisms to enforce them. A policy that states "all AI tools must be reviewed before use" is only as effective as the organization's ability to know when AI tools are being used without review.

The measurement problem is structural. Shadow IT detection in traditional environments works because software assets leave signatures: network connections, installed applications, process names, DNS queries. AI systems used through browsers, embedded in SaaS platforms, or accessed via shared API keys leave signatures that blend into the background noise of normal enterprise activity. An organization that conducts a shadow AI inventory using only traditional discovery methods is measuring the visible fraction of a much larger iceberg.

This creates a governance trap. Security leaders who rely on visibility metrics derived from inadequate discovery methods will underestimate shadow AI exposure, which will inform underinvestment in discovery tooling, which will perpetuate the underestimation. Breaking this cycle requires acknowledging that current visibility metrics are systematically biased downward, and investing in discovery capability specifically designed for AI-era asset profiles.

Industry research on AI agent monitoring provides a concrete example of this dynamic. The CSA found that 82% of enterprise security and IT teams had discovered at least one AI agent or workflow in the past year that they did not previously know existed [6]. The discovery happened after the agent was deployed and operating. In the majority of cases, the agent had been running for some period—accessing data, making decisions, generating outputs—before security teams became aware of its existence.

It bears noting that the quantitative data underlying this analysis draws predominantly from surveys conducted by security vendors, technology research firms, and industry practitioners—sources that have an inherent interest in documenting the scale of the problem their products or research address. The figures cited throughout this whitepaper should be treated as directional indicators of exposure magnitude rather than precise population measurements. Where figures from independent or government sources exist, they tend to corroborate the order of magnitude reported by vendor surveys.

# 6. Building AI Visibility: The Asset Management Imperative

## 6.1 AI Asset Discovery

Effective governance of shadow AI begins with systematic discovery—not the ad hoc discovery that follows an incident, but continuous, structured discovery designed to surface AI assets across all three categories identified in this whitepaper. Discovery for AI requires approaches that traditional shadow IT tooling does not provide.

For standalone AI tools, discovery mechanisms should extend beyond network-level detection to include endpoint telemetry, browser extension analysis, and OAuth token auditing. Browser-based AI tool use is effectively invisible to network detection but leaves traces in browser history, extension permissions, and—where tools request OAuth access—in identity provider token logs. Auditing OAuth token grants across the enterprise identity fabric surfaces AI integrations that users have authorized to access corporate data, regardless of whether IT sanctioned the integration.

For embedded SaaS AI features, discovery requires a different approach: auditing existing SaaS platforms for AI capabilities that are active but were not reviewed. This means maintaining an inventory not of AI tools as a separate category, but of AI capabilities as attributes of all enterprise software assets. Each SaaS platform in the enterprise portfolio should be evaluated periodically for AI features that have been added since the last security review, and those features should trigger a reassessment of the data handling and access implications of the platform.

For AI agents, discovery requires treating agents as first-class identity entities. An agent inventory should record each agent's identity, its data access permissions, the systems it can invoke, the humans who authorized it, and its operational scope. Tools in the AI Trust, Risk, and Security Management (AI TRiSM) category are emerging specifically to address agent discovery and categorization [2], and several established security vendors—including CrowdStrike and Qualys—have extended their platforms to provide AI asset visibility capabilities [15][16].

## 6.2 Risk Classification

Not all AI use poses equal risk, and governance frameworks that treat all shadow AI as equivalent will either become unenforceable through over-restriction or ineffective through under-restriction. Risk classification provides the basis for proportionate governance: different categories of AI use warrant different levels of

control, monitoring, and remediation urgency.

A practical risk classification framework for AI assets should account for at least three dimensions: the sensitivity of data that the AI system can access; the degree of autonomy the system exercises (reactive assistant versus proactive agent); and the existence of human review mechanisms for the system's outputs. A chatbot that an employee uses to draft a non-sensitive internal memo represents a different risk profile than an agent that autonomously processes customer contracts, generates compliance-relevant documentation, and writes to enterprise systems of record.

The EU AI Act's risk tier framework—unacceptable, high, limited, and minimal risk—provides a regulatory anchor for risk classification, though organizations should adapt that framework to their specific risk tolerance and regulatory context. For agentic systems, the CSA's MAESTRO framework provides a seven-layer threat model that enables risk assessment at the level of individual architectural components: the foundation models layer, the data operations layer, the agent frameworks layer, the deployment and infrastructure layer, the evaluation and observability layer, the security and compliance layer, and the agent ecosystem layer [12]. Risk that chains across multiple MAESTRO layers—a threat that begins at the foundation models layer and propagates through agent frameworks to the agent ecosystem layer—warrants the most urgent governance attention.

## 6.3 Continuous Monitoring

Inventory and classification are point-in-time activities; the shadow AI problem is continuous. An enterprise that conducts a thorough AI asset inventory in June 2026 and then repeats the exercise in June 2027 will find the landscape has changed substantially, because the AI tool ecosystem is changing substantially. Continuous monitoring is required to detect new AI asset introductions, changes to the permissions or capabilities of existing AI assets, and behavioral anomalies from AI agents operating within the enterprise environment.

Monitoring for agentic AI should be modeled on the identity monitoring discipline that has developed for human and non-human identities in cloud environments. This means establishing baselines for each agent's normal activity pattern—which APIs it calls, in what frequency, what data it accesses—and alerting on deviations from that baseline. An agent that has been accessing a defined set of SharePoint libraries for six months and suddenly begins querying HR system APIs has exhibited a behavioral change that warrants investigation, regardless of whether that change was intentional or the result of adversarial manipulation.

The CSA's Shadow Access and AI research describes a four-stage lifecycle for AI-assisted identity governance—continuous monitoring, context and visualization, automated risk analysis, and remediation—that provides a practical operational model for this work [9]. The goal is not to eliminate all AI use, but to ensure that the AI use occurring within the enterprise is visible, attributed, and within the bounds of policy.

# 7. Governance Architecture for the Shadow AI Era

## 7.1 Policy That Channels Rather Than Prohibits

The most consistent finding in shadow AI research is that prohibition without alternatives drives AI use to less observable channels. When enterprises provide sanctioned AI tools that meet the productivity needs driving shadow AI use, unauthorized use drops substantially—research suggests by as much as 89% when approved alternatives are accessible and performant [1]. Effective shadow AI governance therefore begins not with a restriction policy but with a sanctioned alternatives strategy: determining which AI use cases are occurring in the workforce, identifying or developing enterprise-approved tools that address those use cases, and making sanctioned alternatives accessible enough that using an approved tool requires less friction than using an unsanctioned one.

Policy must also address the embedded AI dimension that sanctioned alternatives cannot resolve. Enterprises should establish a process for evaluating AI features in existing SaaS platforms, with defined criteria for approval, conditional approval, or disabling. This process should be triggered by vendor AI feature announcements and by periodic platform reviews, not only by incident discovery. Data handling terms for AI features should be scrutinized independently of the underlying platform terms, as AI features often involve materially different data retention, training use, and third-party sharing provisions than the base platform—a pattern documented in the data processing addenda of major enterprise AI platforms including Microsoft 365 Copilot and Salesforce Einstein.

## 7.2 Identity-Centric Controls for AI Agents

Because AI agents interact with enterprise systems through non-human identities, the security controls that apply to those identities are the primary technical mechanism for governing agent behavior. This means applying identity governance disciplines to AI agents: assigning each agent a unique, non-shared identity; defining and documenting the minimum permissions necessary for the agent's intended function; establishing a review process for agent permission grants; and implementing a lifecycle management process that decommissions agent identities when the use case ends.

Just-in-Time (JIT) access provisioning, which grants permissions only for the duration of a specific task rather than maintaining standing access, is particularly well-suited to agentic AI use cases, where a well-scoped agent requires elevated permissions intermittently but not continuously. The CSA's confronting shadow access risks research identifies JIT access as a key mitigation for the over-permissioned identity patterns that enable shadow access exploitation [8]. Applying JIT to AI agent identities limits the blast radius of a compromised or manipulated agent to the time window and scope of a specific permitted task.

## 7.3 Agentic AI Lifecycle Governance

AI agents should be subject to a full software development lifecycle governance process—not treated as configuration items that any team member can create and deploy at will. An agent governance lifecycle includes: a pre-deployment security assessment addressing data access scope, autonomy level, and output verification mechanisms; a formal approval process with defined accountability for the agent's behavior; deployment into a monitored environment with behavioral baselines established before production use; periodic review of the agent's permissions relative to actual use patterns; and a defined decommissioning process that revokes credentials and removes system integrations when the agent is retired.

This governance model is not unprecedented. It is the model that organizations with formal software development lifecycle and vendor risk management programs already apply to software systems, third-party integrations, and cloud workloads. The challenge is extending that model to a category of assets that most organizations do not yet track in their configuration management databases and that most security teams do not yet have operational processes to govern. The investment in extending existing governance disciplines to AI agents is substantially lower than building new disciplines from scratch, and in most cases substantially lower than the cost of an AI-related security incident—which industry data suggests can add $670,000 or more to breach costs [17].

# 8. CSA Resource Alignment

The Cloud Security Alliance has developed an extensive body of research directly applicable to the shadow AI visibility and governance challenge. Security practitioners addressing this problem should draw on these resources as a coordinated framework rather than independent publications.

The **AI Controls Matrix (AICM)** provides the foundational control framework for organizations governing AI systems. With 243 control objectives across 18 security domains, the AICM is a superset of the Cloud Controls Matrix that extends traditional cloud security controls to address AI-specific risks including model security, data governance for AI, and AI supply chain integrity [10]. The AICM's control domains for AI asset management, AI inventory, and AI access control are directly applicable to the shadow AI visibility problem, providing the control structure within which discovery, classification, and monitoring processes should be designed.

The **MAESTRO Agentic AI Threat Modeling Framework** provides the architectural vocabulary for understanding where shadow AI risks materialize within complex agentic systems [12]. MAESTRO's seven-layer model—from foundation models through the agent ecosystem—enables security teams to reason about how threats chain across layers: a prompt injection at the foundation models layer that influences agent frameworks behavior, which in turn creates unauthorized access at the agent ecosystem layer. Applying MAESTRO to shadow AI scenarios, particularly those involving autonomous agents with broad system access, helps security architects identify the highest-consequence risk chains and prioritize controls accordingly.

The **Shadow Access research series**—comprising "Defining Shadow Access," "Confronting Shadow Access Risks," and "Shadow Access and AI"—provides essential grounding for the identity governance dimension of the shadow AI problem [7][8][9]. These publications define the shadow access concept as unmonitored, unauthorized, invisible, unsafe, and generally over-permissioned access; trace how AI systems both create and amplify shadow access risks; and describe the four-stage lifecycle model for AI-assisted shadow access detection and remediation. Organizations building AI agent governance programs should treat the shadow access series as required reading, as the non-human identity governance challenges that shadow AI creates are a direct extension of the shadow access problem these publications address.

The **AI Organizational Responsibilities** series addresses the governance, accountability, and organizational design dimensions of AI security. The third volume, focused on AI Tools and Applications, is particularly relevant to shadow AI governance, as it addresses the RACI framework for AI procurement and approval, the supply chain security obligations that apply to third-party AI tools, and the organizational structures needed to sustain ongoing AI security governance [11]. These guidance documents establish the accountability architecture within which shadow AI discovery and control programs must operate.

The **STAR for AI** program provides a mechanism for verifying AI vendor security postures through standardized self-assessment and third-party attestation. Organizations addressing the embedded SaaS AI problem—determining which AI features in existing vendor platforms are adequately secured—can use the STAR for AI framework to structure vendor assessments and to require consistent security information from AI platform providers as a condition of AI feature approval.

Finally, the CSA's **Zero Trust guidance** provides the access control philosophy that should govern AI agent identity management. The principle of never trust, always verify—instantiated through least-privilege access, continuous authentication, and micro-segmentation—applies directly to AI agents, which should be extended precisely and no more access than their current, specific task requires. Zero Trust architectures that already enforce these principles for human and traditional non-human identities should be extended explicitly to cover AI agent identities, treating agents as a new category of identity subject to the same verification and access control discipline.

# 9. Conclusions and Recommendations

The shadow AI visibility crisis is not a temporary condition that will resolve as AI tools mature. It is a structural feature of a technology adoption wave that has, once again, outpaced the governance architecture built to manage it. The shadow AI problem is solvable with the same disciplines that enterprise security has applied to cloud services, SaaS applications, and non-human identities—applied specifically to AI. The challenge is the speed and breadth of the adoption that has occurred without those disciplines in place.

Organizations should treat the following as near-term priorities:

**Establish AI asset visibility as a security function.** AI discovery, inventory, and classification should be explicitly owned by the security organization and resourced accordingly. Existing shadow IT discovery tooling should be evaluated for AI coverage gaps, and supplementary tooling—browser extension analysis, OAuth token auditing, network AI traffic detection, AI TRiSM platforms—should be assessed and prioritized based on the organization's shadow AI risk profile.

**Audit existing SaaS platforms for AI features that have not been reviewed.** Every SaaS vendor with an active enterprise contract should be evaluated for AI features added since the last security review. Features with access to sensitive data categories should be subject to a risk assessment that addresses AI-specific data handling, training use terms, and prompt injection vulnerability. Features that fail this assessment should be disabled until they can be approved through the standard procurement process.

**Treat AI agents as first-class identity entities.** Every AI agent operating in the enterprise environment should have a documented identity, defined minimum-necessary permissions, a designated human owner accountable for its behavior, and a place in the organization's identity governance processes. Shared API keys as the primary authentication mechanism for AI agents should be replaced with scoped, auditable credentials subject to regular rotation and access review.

**Provide sanctioned AI alternatives for common use cases.** Understanding where shadow AI use is occurring is the first step to channeling it toward sanctioned alternatives. This does not require enterprise development of AI tools—it requires understanding the productivity needs that are driving shadow AI use, and ensuring that approved tools exist to meet those needs, with sufficient accessibility and performance that friction does not push users toward unsanctioned alternatives.

**Engage with the regulatory calendar.** The EU AI Act's enforcement milestones, sector-specific AI regulations, and evolving data protection requirements create a governance compliance timeline that cannot be met without the AI asset inventory that shadow AI visibility provides. Organizations should treat regulatory deadlines as external forcing functions for the governance investments that security risk alone should motivate.

The foundational insight of this whitepaper is simple: every AI security control—access control, data governance, behavioral monitoring, incident response—is predicated on knowing that the AI system exists. Shadow AI is, at its core, a visibility deficit, and the security controls that cannot be applied to systems that cannot be seen will remain aspirational until that visibility is achieved. The frameworks, tooling, and operational processes to achieve it exist. The question is whether organizations will apply them before the next significant shadow AI incident makes the cost of not applying them undeniable.

# References


[1] Unseen Security. "The State of Shadow AI 2026." Unseen Security, 2026.

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

[3] Gartner. "Gartner Identifies Six Steps to Manage AI Agent Sprawl." Gartner Press Release, April 28, 2026.

[4] Gravitee. "State of AI Agent Security 2026 Report: When Adoption Outpaces Control." Gravitee, 2026.

[5] Cloud Security Alliance. "Shadow AI Agents: The Insider Threat You're Not Monitoring Yet." CSA Blog, May 26, 2026.

[6] Cloud Security Alliance. "The Shadow AI Agent Problem in Enterprise Environments." CSA Blog, April 28, 2026.

[7] Cloud Security Alliance. "Defining Shadow Access: The Emerging IAM Security Challenge." CSA, 2023.

[8] Cloud Security Alliance. "Confronting Shadow Access Risks: Considerations for Zero Trust and Artificial Intelligence Deployments." CSA, 2024.

[9] Cloud Security Alliance. "Shadow Access and AI." CSA, 2024.

[10] Cloud Security Alliance. "AI Controls Matrix." CSA, 2025.

[11] Cloud Security Alliance. "AI Organizational Responsibilities: AI Tools and Applications." CSA, 2025.

[12] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[13] BlackFog. "CamoLeak: How GitHub Copilot Became an Exfiltration Channel." BlackFog, April 8, 2026.

[14] TrustLogix. "AI Agent Sprawl Is the New Shadow IT." TrustLogix, April 22, 2026.

[15] Qualys. "From Shadow Models to Audit-Ready AI Security: A Practical Path with Qualys TotalAI." Qualys Blog, March 10, 2026.

[16] CrowdStrike. "New CrowdStrike Innovations Secure AI Agents and Govern Shadow AI." CrowdStrike Blog, March 23, 2026.

[17] Vectra AI. "Shadow AI Explained: Risks, Costs, and Enterprise Governance." Vectra AI.

[18] DSALTA. "Shadow AI Compliance: Risks, Governance & 2026 Guide." DSALTA, May 11, 2026.

[19] Second Talent. "Top 50 Shadow AI Statistics 2026." Second Talent, 2026.

[20] Cloud Security Alliance. "The State of AI Security and Governance." CSA/Google Cloud, December 2025.