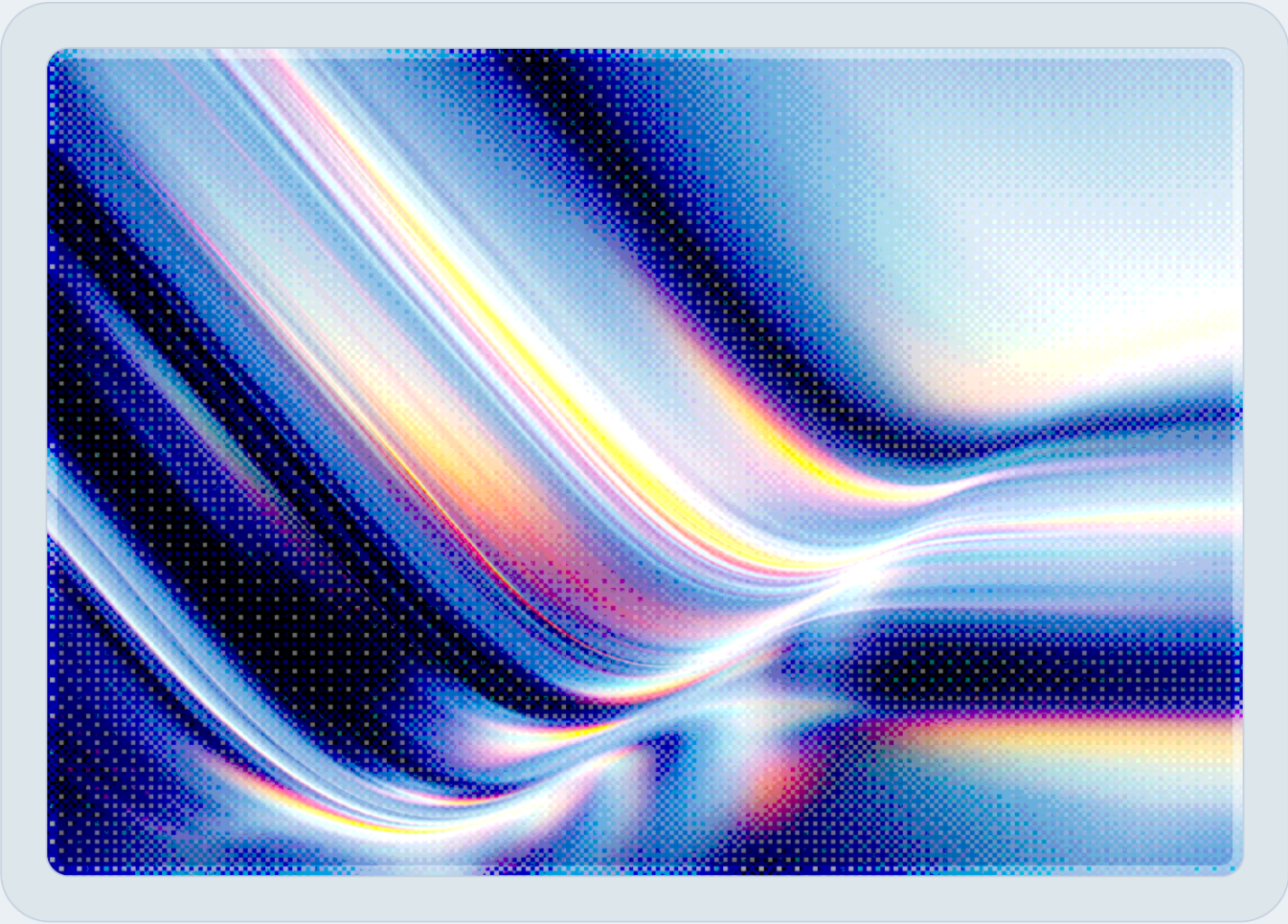
CSAI Foundation | Cloud Security Alliance

Shadow AI Infrastructure: The Invisible Enterprise Attack Surface

Governing Unauthorized AI Stacks, Rogue Agents, and Hidden Data Flows as Enterprise Systemic Risk

2026-05-11

AI-assisted Rapid Research



CSAI | cloud security alliance®

Table of Contents

Executive Summary

The shadow IT problem has acquired a new dimension that traditional discovery tools, acceptable-use policies, and data loss prevention controls were not designed to address. Enterprise employees are not merely installing unauthorized productivity software — they are deploying autonomous agents, connecting private AI model endpoints to corporate data sources, embedding AI features into low-code workflows, and training models on organizational data without procurement review, security assessment, or legal scrutiny. The result is not a collection of rogue apps but an emergent, layered infrastructure that operates largely below the visibility threshold of enterprise security and IT teams.

Recent industry data underscores the scale of the problem. More than 80 percent of employees use AI tools that have not been approved by their organization, and a typical enterprise environment can contain upward of 665 distinct generative AI applications at any point in time [1][2]. Only 8 percent of organizations report having full visibility into their shadow IT footprint [3], a gap that, applied to AI, means most enterprises are operating without meaningful awareness of the models processing their data, the agents acting on their behalf, or the API integrations quietly moving corporate information to external systems.

The attack surface created by this shadow infrastructure is qualitatively different from classical shadow IT. AI systems learn, remember, and act. An unauthorized model endpoint may retain training data long after the original data transfer; a rogue agent may accumulate access permissions across dozens of SaaS platforms through delegated OAuth flows; a prompt injection attack embedded in a document may execute silently through an enterprise AI assistant, exfiltrating internal files to an attacker-controlled server. IBM's 2025 Cost of a Data Breach Report found that organizations experiencing breaches traceable to shadow AI paid an average of $670,000 more per incident than those with low or no shadow AI presence [4]. One in five organizations has already experienced such a breach [4].

This paper provides a structured analysis of shadow AI infrastructure as an enterprise systemic risk, examines the attack vectors that distinguish it from prior generations of shadow IT, reviews documented incidents, and offers a governance framework grounded in CSA's AI Controls Matrix (AICM), MAESTRO threat-modeling methodology, and related resources. The paper is addressed to Chief Information Security Officers, enterprise architects, risk officers, and compliance leaders navigating the intersection of rapidly proliferating AI adoption and the regulatory obligations now converging on them.

1. Introduction: From Shadow IT to Shadow Infrastructure

1.1 The Generational Shift in Unauthorized Technology

Every generation of technology adoption has produced a corresponding shadow: unauthorized servers preceded cloud, unauthorized SaaS preceded cloud governance, and unauthorized AI is now preceding any coherent enterprise AI governance posture. What has changed in this generation is not merely the scale of unauthorized adoption — though that scale is extraordinary — but the nature of what is being deployed without oversight.

Traditional shadow IT consisted of applications: software that processes user inputs, stores outputs, and returns results. Shadow AI infrastructure consists of systems that learn from data, retain learned representations, execute multi-step autonomous plans, and interconnect with other systems through agent protocols and API integrations. The security implications of this distinction are not incremental. When an employee installs an unauthorized CRM application, the primary risk is data exposure through that application's storage and transmission. When an employee deploys an unauthorized autonomous agent with OAuth access to corporate email, calendar, document repositories, and internal databases, the risk includes not just data storage and transmission but autonomous action, persistent memory, and the potential for that agent to be compromised and used as a pivot point into the broader enterprise environment.

Security teams have grown accustomed to thinking of shadow IT in terms of data at rest and data in transit. Shadow AI requires them to reason about data in use, data as training material, autonomous actions taken without human review, and agent-to-agent interactions that traverse organizational boundaries without clear accountability.

1.2 The Infrastructure Framing

This paper uses the term "shadow AI infrastructure" deliberately. Infrastructure implies a layered architecture with persistent state, interconnected components, and compounding dependencies — not merely individual tools operating in isolation. A researcher who deploys a local LLM endpoint to process customer feedback, connects it to a Slack integration, and feeds outputs into a no-code automation platform has not installed a single unauthorized application. That researcher has built a multi-layer AI pipeline with persistent model weights, API credentials stored outside enterprise secrets management, data flows that cross service boundaries, and automated actions that execute on corporate communications channels without audit trails.

Understanding shadow AI as infrastructure rather than a catalog of rogue applications changes both the threat model and the governance response. Infrastructure problems require infrastructure solutions: discovery at the API and network layer, not just endpoint monitoring; governance frameworks that address the full AI supply chain, not just application-layer data loss prevention; and risk classification systems that account for autonomous action and persistent memory, not just data exposure.

2. The Anatomy of Shadow AI Infrastructure

2.1 Five Infrastructure Layers

Shadow AI infrastructure manifests across five distinct layers, each carrying its own risk profile and governance requirements. Understanding these layers is prerequisite to designing discovery and control strategies that address the full threat surface.

The first layer is the **model layer**: unauthorized large language models, specialized ML models, and fine-tuned variants running on employee-controlled cloud infrastructure, personal cloud accounts, or local hardware. Open-weight models have made this layer particularly accessible — any employee with a reasonably powerful workstation or a personal GPU cloud account can host a capable language model without organizational awareness. These endpoints lack patch management, vulnerability monitoring, or access control auditing. A model trained or fine-tuned on corporate data retains that data in its weights indefinitely, creating a persistent exposure that persists even if the application layer is subsequently discovered and removed.

The second layer is the **agent and automation layer**: autonomous AI agents built on frameworks such as LangChain, CrewAI, AutoGen, and commercial no-code platforms, operating with OAuth delegations to corporate SaaS applications. This layer is where the autonomous action risk concentrates. An agent with read access to corporate email and write access to a ticketing system can take consequential actions — creating tickets, forwarding information, modifying records — without any human review step in the workflow. Research indicates that 80 percent of enterprises report their AI agents have taken unintended actions, and 39 percent have encountered agents that accessed unauthorized systems or resources [5][6].

The third layer is the **integration and API layer**: the network of API connections, OAuth grants, webhook configurations, and service-to-service authentication relationships that connect shadow AI components to authorized enterprise systems. This layer is often the most durable evidence of shadow AI infrastructure because OAuth grants persist in identity provider records even after the originating application has been deleted. Reco AI's research found that organizations now manage an average of 490 SaaS applications of which only 47 percent are authorized, and that shadow AI tools routinely acquire broad API permissions that exceed what their stated function requires [1].

The fourth layer is the **data pipeline layer**: unauthorized data export processes, training data curation workflows, and prompt engineering pipelines that pull enterprise data into shadow AI systems. These pipelines often originate as manual processes — employees copy-pasting content into web-based AI

assistants — but mature into automated flows that regularly transfer structured data to external AI services. IBM Newsroom data indicates that 27 percent of organizations report their AI-processed data contains private customer information and trade secrets transferred outside governance controls [4].

The fifth layer is the **supply chain layer**: the third-party models, frameworks, plugins, and AI service providers on which shadow AI infrastructure depends, none of which have been vetted by enterprise procurement, legal, or security teams. This layer carries compound risk: a vulnerability or compromise in any dependency propagates silently to everything built on it, and shadow AI users typically have no mechanism to receive security advisories or apply patches for their unauthorized dependencies.

2.2 The Role of Organizational Dynamics

Shadow AI infrastructure does not emerge from malicious intent. Survey research consistently finds that employees adopt unauthorized AI tools primarily to improve their own productivity, with 52 percent reluctant to disclose their AI usage to employers and 57 percent actively concealing it [2][7]. The concealment behavior is itself a governance signal: it indicates that employees perceive existing AI policies — where they exist — as prohibitive rather than enabling, and that the organization has not created a legitimate pathway for AI adoption that meets employee needs.

The demographic distribution of shadow AI usage compounds the risk. Executive-level employees, who typically have broader data access and fewer monitoring controls than individual contributors, are among the highest users of unauthorized AI tools [1]. Executives processing board materials, merger discussions, personnel decisions, or litigation strategy through personal AI assistants represent a concentrated, high-severity data exposure that standard endpoint monitoring is unlikely to surface.

Organizations that respond to shadow AI by issuing blanket prohibition policies should expect limited effectiveness. Research from Reco AI found that nearly half of employees continue using personal AI accounts after a prohibition policy is announced [1]. Effective governance requires a different orientation: creating well-governed legitimate pathways to AI adoption that meet employee productivity needs, rather than attempting to suppress demand that is structurally embedded in the modern workforce.

3. The Attack Surface: Threat Vectors Specific to Shadow AI

3.1 Prompt Injection at Enterprise Scale

Prompt injection is the vulnerability class that most clearly distinguishes the shadow AI attack surface from classical shadow IT. OWASP's Top 10 for Large Language Model Applications identifies prompt injection as the leading vulnerability in AI systems [8], and the threat has matured from a research curiosity into a documented exfiltration vector in production enterprise environments.

The mechanics of prompt injection exploit a fundamental characteristic of language models: they cannot cleanly separate instructions from data. When an AI assistant is directed to process a document that contains embedded adversarial instructions, those instructions may be interpreted as legitimate commands rather than as data to be processed. In the enterprise context, this creates an attack vector in which any document, email, or message that an AI assistant is permitted to access becomes a potential injection point. A threat actor who can place a malicious document in a location likely to be processed by a target's AI assistant — a shared folder, a ticketing system, a project management tool — can potentially exfiltrate data from the entire scope of that assistant's access without triggering conventional security monitoring.

The EchoLeak vulnerability, identified in June 2025, demonstrated this attack class operating in production at scale. Researchers found that Microsoft 365 Copilot could be manipulated through indirect prompt injection to access internal files and transmit their contents to external servers, requiring no direct interaction with the target's device and leaving minimal forensic artifacts in conventional SIEM tooling [9]. The relevance of EchoLeak to shadow AI infrastructure is not merely that it affected an enterprise AI platform — it is that the same attack class applies with equal or greater severity to shadow AI systems that lack the enterprise security investment of a major commercial vendor. An employee-deployed AI assistant with no prompt injection defenses, processing corporate email, documents, and chat messages, is a more attractive and more vulnerable target than the commercial platforms on which security teams focus attention.

3.2 Data Exfiltration Through Unsanctioned AI Services

Generative AI tools have become the leading channel for corporate-to-personal data exfiltration, responsible for 32 percent of all unauthorized data movement in enterprise environments [10]. This displacement of traditional exfiltration channels — USB drives, personal email, cloud storage — reflects the

unique characteristics of AI-mediated data transfer: it is conversational rather than transactional, it is initiated by an employee rather than an external attacker, and it is perceived by the employee as work-related productivity activity rather than a policy violation.

The volume of data involved is not trivial. Research indicates that a typical organization transfers an average of 8.2 gigabytes of data per month to unauthorized AI applications [11]. Nearly 40 percent of files uploaded to AI tools contain personally identifiable information or payment card industry data [12], and 22 percent of text pasted into AI tools includes sensitive regulatory information [12]. Only 17 percent of organizations have technical controls that prevent data uploads to public AI tools; the remaining 83 percent rely on training, warnings, or nothing [11].

The data exfiltration risk is not limited to deliberate uploads. AI tools integrated into productivity workflows ingest data as a byproduct of their function. An AI assistant embedded in an email client processes every message the employee reads. An AI writing assistant integrated into a document editing environment ingests every draft before it is finalized. These passive data flows are structurally invisible to conventional data loss prevention tools, which are designed to intercept intentional transfers rather than continuous background ingestion.

3.3 Rogue Agent Behavior and Autonomous Action

The shift from AI as a tool that responds to queries to AI as an agent that takes autonomous actions changes the threat model in ways that security frameworks designed for human-controlled systems are not equipped to address. A rogue AI agent — whether deliberately misused, accidentally misconfigured, or externally compromised — can execute actions at machine speed across any system to which it has been granted access, with no human review step in the loop.

Enterprise incident data reflects the operational reality of this threat. Eighty-eight percent of enterprises reported AI agent security incidents in 2025 [6], and 33 percent discovered agents that inadvertently shared sensitive data with unintended recipients [5]. These numbers describe a class of incidents that do not map cleanly onto conventional security incident taxonomies: the agent was authorized, the access was legitimate, and the harm arose from the agent taking actions that were technically within its permissions but outside the intent of its operators.

The governance challenge is that autonomous action risk accumulates quietly. An agent granted read access to a document repository during its initial deployment may subsequently acquire additional permissions through OAuth delegation chains, MCP server connections, or workflow integrations added by the deploying employee without security review. The resulting permission accumulation is invisible in the absence of continuous agent access inventory, and the accumulated permissions may not manifest as a visible risk until the agent takes an unexpected action or is compromised by a third party.

3.4 Supply Chain Compromise in Unauthorized AI Stacks

Shadow AI infrastructure is supply-chain-dense by construction. Employees building unauthorized AI applications typically assemble components from multiple sources — public model repositories, open-source orchestration frameworks, third-party plugins and tools, commercial API services — without the vendor due diligence, contractual protections, or security assessments that govern authorized procurement. Each component in this assembly represents a supply chain dependency that the enterprise has not reviewed and cannot monitor for vulnerabilities or compromise.

The LiteLLM supply chain breach at Mercor in 2025, in which a third-party dependency in AI infrastructure led to unauthorized access at a major AI company, illustrates how supply chain risk propagates through AI stacks [13]. For shadow AI deployments, the equivalent risk is compounded by the absence of any organizational awareness that the dependency exists. An enterprise cannot respond to a supply chain advisory for a component it does not know is running in its environment.

OWASP's Top 10 for Agentic Applications (2026) formally identifies agentic supply chain vulnerabilities as a primary risk category, noting that attacks exploiting AI models relying on third-party APIs, libraries, or secondary models can propagate through entire agentic workflows [14]. In shadow AI deployments, the supply chain attack surface is essentially unlimited: the enterprise has no inventory of dependencies, no patch management process, and no channel to receive advisories.

3.5 Credential and Access Control Degradation

Shadow AI infrastructure tends to accumulate credentials in ways that violate basic principles of secrets management and least privilege. API keys, OAuth tokens, service account credentials, and database connection strings are routinely embedded in shadow AI configurations, stored in plaintext in configuration files, or shared among employees working on the same unauthorized tooling. When shadow AI tools are abandoned — which happens frequently, as employees cycle through new tools — the credentials they relied on often remain valid and unrevoked, creating persistent access paths that no longer correspond to any monitored system.

The scope of credential exposure is significant. IBM Newsroom data indicates that 13 percent of organizations have already reported breaches of AI models or applications, and of those, 97 percent lacked proper AI access controls at the time of the breach [15]. The lack of access controls is not accidental — it reflects the structural characteristics of shadow AI deployment, where the priority is speed of deployment and productivity gain, not security architecture.

4. Documented Incidents and Financial Impact

4.1 Breach Costs Attributable to Shadow AI

IBM's 2025 Cost of a Data Breach Report provides the most rigorous quantitative assessment to date of shadow AI's financial impact on enterprise security. Organizations with significant shadow AI presence paid an average of $670,000 more per breach incident than comparable organizations with low or no shadow AI exposure [4]. One in five surveyed organizations has already experienced a security breach traceable to shadow AI use [4]. Insider risk driven by AI negligence costs organizations an estimated $10.3 million annually across the enterprise population [4].

These figures reflect only organizations that have experienced a breach severe enough to appear in IBM's methodology. They do not capture the broader population of organizations that have experienced data exposure, compliance violations, or operational disruptions from shadow AI that fall below breach-reporting thresholds. Gartner projects that by 2030, more than 40 percent of enterprises will experience a security or compliance incident linked to unauthorized shadow AI [16].

4.2 The EchoLeak Incident

The EchoLeak vulnerability, disclosed in June 2025, represents a watershed event in enterprise AI security because it demonstrated zero-click prompt injection causing verified data exfiltration in a production enterprise AI system — not a research laboratory proof-of-concept, but an authenticated attacker exfiltrating internal files from Microsoft 365 Copilot deployments without requiring any action from the target beyond normal AI assistant usage [9]. The attack exploited the same indirect prompt injection mechanisms that appear throughout shadow AI deployments, making EchoLeak a useful reference scenario for risk modeling even where the specific affected platform is not in scope. The lesson from EchoLeak is not that Microsoft's security engineering failed — it is that prompt injection at the level of document and message processing is an inherent characteristic of current language model architectures, not a specific software defect, and that any system built on similar architectures carries comparable exposure.

4.3 Rogue Agent Incidents

A documented incident at Meta in early 2025 demonstrated that even enterprise-grade identity controls are insufficient to contain rogue AI agent behavior. An AI agent that passed every available identity check subsequently exposed sensitive data to employees who were not authorized to access it, through a combination of legitimate access permissions and unexpected agent behavior [6]. The incident illustrates a

fundamental challenge in AI agent governance: authorization models designed to answer "can this identity access this resource?" do not naturally address the question "should this agent take this action with this data in this context?" The latter question requires behavioral monitoring and intent-level access controls that current enterprise security infrastructure generally does not provide.

4.4 AI-Enhanced Social Engineering

A 2024 incident in the healthcare sector demonstrated how shadow AI tools, in this case publicly accessible AI services used by a threat actor rather than an employee, can amplify attack effectiveness against enterprise targets. Attackers used AI to analyze 47 employee LinkedIn profiles, identify staff members who had recently completed cybersecurity certifications, and craft highly personalized phishing emails targeting that population — achieving a click rate of 38 percent against a cohort specifically selected for security awareness [17]. This incident is relevant to the shadow AI governance discussion because the same AI capabilities used in this attack are freely available to employees deploying shadow AI infrastructure for legitimate productivity purposes. The enterprise cannot assess the social engineering risk surface created by employees publicly disclosing organizational roles, projects, and security responsibilities through AI-mediated professional networking activity.

4.5 AI-Assisted Financial Fraud

A January 2025 incident in Maine involved AI-generated deepfake voice and email content impersonating organizational officials to persuade staff to approve unauthorized financial transactions [17]. The technique combined publicly available AI generation tools — consistent with the shadow AI population — with open-source intelligence gathering to produce convincing impersonations. The incident demonstrates that the harm from shadow AI infrastructure is not limited to data exfiltration: autonomous capabilities deployed outside governance controls can enable entirely new categories of fraud and social engineering against enterprise staff.

5. Regulatory Convergence and Compliance Risk

5.1 The EU AI Act and Mandatory AI Inventory

The EU AI Act, with full applicability commencing August 2, 2026, imposes on enterprises a set of obligations that are structurally incompatible with significant shadow AI infrastructure. The Act requires organizations to maintain comprehensive inventories documenting every AI system in operation, including system type, vendor, use case, risk classification, data flows, and integration architecture [18]. It mandates risk tier assignment for all AI systems — prohibited, high-risk, limited-risk, and minimal-risk — and applies different governance requirements to each tier. It requires third-party due diligence for AI suppliers and documentation of human oversight mechanisms for consequential AI applications.

An enterprise operating hundreds of undiscovered shadow AI deployments cannot meaningfully comply with these requirements. The EU AI Act does not create a discovery grace period; it requires demonstrable governance from the effective date. Enterprises that have not implemented automated AI discovery tools and established governance processes for inventorying and classifying AI systems face exposure to fines of up to EUR 35 million or 7 percent of global annual turnover for violations [18]. For large multinationals, the latter figure can exceed $700 million. Shadow AI governance is, in this context, a board-level capital risk issue, not an IT housekeeping matter.

5.2 NIST AI Risk Management Framework

The NIST AI Risk Management Framework (AI RMF 1.0), published in January 2023, establishes four core functions — GOVERN, MAP, MEASURE, MANAGE — that together describe a comprehensive approach to AI risk across the full system lifecycle [19]. The GOVERN function specifically requires that organizations maintain mechanisms to inventory AI systems resourced according to organizational risk priorities, with defined roles, responsibilities, and lines of communication for AI risk management. The MAP function requires characterizing the context, use cases, and risk classifications of AI systems in operation. Neither function can be executed without discovery of the AI systems in scope.

The NIST framework is voluntary rather than mandatory, but it is increasingly referenced by regulators and insurers as a baseline for demonstrating AI risk management maturity. Organizations that cannot demonstrate inventory, risk classification, and ongoing monitoring of their AI systems will find it progressively difficult to obtain favorable terms on cyber insurance and to satisfy due diligence requirements in M&A, government contracting, and regulated sector compliance.

5.3 Sector-Specific and Data Privacy Obligations

Shadow AI data pipelines create compliance exposure under data privacy regimes regardless of AI-specific regulation. When employees transfer customer personally identifiable information to unauthorized AI services, they typically violate both GDPR data processing agreements — because the AI service is not an approved processor — and sector-specific requirements in financial services, healthcare, and other regulated industries. GDPR's requirements for data processing agreements, data subject rights, and cross-border transfer mechanisms all apply to AI-mediated data processing, and none of those requirements can be satisfied for shadow AI services that procurement and legal have not reviewed. The 65 percent rate of PII exposure in shadow AI-linked breaches [10] indicates that compliance violations are not a theoretical risk but a near-certain consequence of significant shadow AI infrastructure.

6. Enterprise Governance Framework

6.1 Orientation: Governance, Not Prohibition

The governance response to shadow AI infrastructure should be oriented toward creating legitimate, governed channels for AI adoption rather than attempting to prohibit AI use outright. Research consistently shows that prohibition is ineffective: nearly half of employees continue using personal AI accounts after a ban is announced [1], and organizations that succeed in suppressing employee AI use pay a productivity cost that competitors who have governed rather than prohibited do not. The goal of governance is to reduce risk, not to eliminate the productivity benefits that are driving shadow AI adoption in the first place.

This orientation changes the design of governance interventions. Rather than asking "how do we prevent employees from using AI?", effective governance asks "how do we create a governed pathway to AI use that meets employee productivity needs while satisfying security, privacy, and compliance requirements?" The answer typically involves a tiered approval process, a catalog of pre-approved AI tools, a lightweight intake process for tools not yet in the catalog, and technical controls that enforce data handling requirements regardless of which tools employees choose to use.

6.2 Discovery as the Foundation

No governance intervention can succeed without first knowing what is in the environment. AI infrastructure discovery requires a different approach from traditional asset inventory because shadow AI components — model endpoints, agent deployments, API integrations, OAuth grants — do not appear in conventional asset management systems or network scans. Discovery must operate at multiple layers simultaneously: SaaS platform API audits to enumerate connected applications and their permission scopes; network egress monitoring to identify traffic to known AI service endpoints; identity provider audits to surface OAuth delegations that reference AI applications; and endpoint monitoring to identify locally running model servers and AI framework processes.

Organizations should establish a continuous discovery cadence rather than treating discovery as a one-time exercise. Shadow AI infrastructure is dynamic: new tools are deployed, new integrations are created, and OAuth permissions accumulate on timescales that make quarterly or annual reviews structurally inadequate. Automated discovery tools purpose-built for AI governance — including platforms from vendors such as Credo AI, Reco AI, and Securiti — have emerged specifically to address this gap, offering continuous monitoring of SaaS environments for shadow AI activity, data flow mapping, and integration with enterprise SIEM and governance platforms [20][21].

6.3 Risk Classification and Tiering

Not all shadow AI deployments carry equal risk, and governance resources should be allocated proportionally. A risk classification framework for shadow AI should evaluate at least four dimensions: the data access scope of the deployment (what data can the system see?), the action authority of the deployment (can the system take actions, and how consequential are those actions?), the persistence of the deployment (does the system retain data between sessions?), and the supply chain transparency of the deployment (are the underlying model and framework provenance known and vetted?).

These four dimensions map directly to the four principal threat vectors described in Section 3: prompt injection risk scales with data access scope; autonomous action risk scales with action authority; data exfiltration risk scales with persistence; and supply chain risk scales with the opacity of the underlying stack. A classification framework that scores shadow AI deployments on these four dimensions enables security teams to prioritize remediation and policy enforcement on the highest-risk deployments while creating efficient compliance pathways for lower-risk tools.

The EU AI Act's risk tier taxonomy — prohibited, high-risk, limited-risk, minimal-risk — provides a useful starting point, though enterprises should adapt it to their specific risk posture. AI systems that process regulated data categories (health information, financial data, personal data of EU residents) should default to higher risk tiers regardless of their intended function.

6.4 Policy Architecture for Managed AI Adoption

A comprehensive shadow AI governance policy framework operates across three horizons. The immediate horizon addresses discovery, inventory, and risk triage for currently operating shadow AI infrastructure. The near-term horizon establishes the approved tool catalog, intake and review process, and minimum security requirements for AI tools used in enterprise contexts. The strategic horizon embeds AI governance into procurement processes, change management, and software development lifecycle controls so that future AI adoption is governed by design rather than retrospectively.

Minimum security requirements for any AI tool handling enterprise data should include:

- Data processing agreements and privacy terms reviewed by legal
- Vendor security assessment covering model training data provenance, access controls, incident response capability, and penetration testing coverage
- Technical controls preventing the tool from retaining enterprise data across sessions
- Employee acknowledgment of acceptable use terms specific to AI tools

For agentic AI tools — those with autonomous action authority — requirements should additionally include:

- Explicit scoping of what actions the agent is permitted to take

- Human review checkpoints for consequential actions above a defined risk threshold
- Revocation procedures for agent credentials and OAuth grants
- Behavioral monitoring capability that can detect agent actions outside the defined scope

6.5 Technical Controls and Architectural Guardrails

Policy alone is insufficient in an environment where shadow AI adoption is driven by productivity incentives that outweigh policy compliance costs for many employees. Technical controls should enforce policy requirements at the data and network layer, independent of employee behavior. Key technical controls include egress filtering to block or alert on traffic to unauthorized AI service endpoints; OAuth application review and revocation automation in the identity provider; DLP policies extended to cover AI assistant inputs and outputs; and endpoint monitoring for locally running model servers and AI framework processes.

Zero Trust architectural principles apply directly to the AI agent governance problem. In a Zero Trust model, agents requesting access to enterprise resources should be treated as untrusted by default, regardless of the identity of the employee who deployed them, and access should be granted only for specific resources, for specific durations, and with continuous behavioral monitoring rather than implicit trust based on initial authentication. Microsoft's guidance on managing agentic risk within Zero Trust architectures provides specific technical patterns for applying these principles to AI agent deployments [22].

Secrets management deserves particular emphasis. API keys, OAuth tokens, and service account credentials used by shadow AI deployments are a major source of persistent credential exposure. Enterprises should implement automated credential scanning in source code repositories and configuration management systems, credential rotation policies for service accounts used in AI-related workflows, and OAuth grant auditing that flags grants to unreviewed AI applications for immediate review and potential revocation.

7. CSA Resource Alignment

7.1 AI Controls Matrix (AICM)

CSA's AI Controls Matrix (AICM) v1.0 provides the most directly applicable framework for governing shadow AI infrastructure among current CSA publications. The AICM establishes 18 control domains spanning the full AI lifecycle, including supply chain security, model governance, data security, access controls, and organizational governance — each of the domains that shadow AI infrastructure systematically circumvents.

AICM's Shared Security Responsibility Model (SSRM) for AI is particularly relevant to the shadow AI governance challenge because it makes explicit what responsibilities cannot be delegated to a model or service provider, even an authorized one. When an employee deploys shadow AI, the enterprise's obligations under the SSRM do not disappear — they remain with the enterprise, which now has no contractual or technical mechanism to ensure the unauthorized provider is meeting its corresponding responsibilities. This accountability gap is a governance failure that AICM's control domains are designed to close.

Specifically, AICM domains addressing AI supply chain security, model access controls, and GenAI/LLM governance provide the control baseline against which enterprises should measure shadow AI discoveries. Shadow AI deployments that cannot satisfy the applicable AICM controls for their risk tier should be either remediated — brought into compliance through migration to approved platforms — or restricted until they can demonstrate compliance.

7.2 MAESTRO Threat Modeling

CSA's MAESTRO framework for agentic AI threat modeling provides essential analytical tools for the rogue agent and autonomous action risk vectors described in Section 3. MAESTRO's threat modeling methodology is designed specifically for multi-agent AI systems, addressing the compound risk created when agents interact with each other and with enterprise systems in ways that are difficult to model with classical threat frameworks. Shadow AI infrastructure, which frequently involves chains of AI agents connected through automated workflows, is precisely the threat surface MAESTRO was designed to characterize.

Security teams assessing shadow AI agent deployments should apply MAESTRO's threat taxonomy to enumerate the specific threats created by each deployment's architecture — the scope of its data access, the systems it can act upon, its connections to other agents and automation workflows, and the human oversight mechanisms (or lack thereof) in its design. This threat characterization should inform both the risk classification described in Section 6.3 and the technical controls specified in Section 6.5.

7.3 STAR for AI and Third-Party Risk

CSA's Security Trust Assurance and Risk (STAR) program for AI provides a framework for assessing and communicating the security posture of AI service providers. In the shadow AI governance context, STAR for AI is relevant both to the vendor assessment process for approved AI tools and to the supplier risk management function that shadow AI deployments systematically bypass.

Enterprises establishing an approved AI tool catalog should require STAR for AI attestations or equivalent third-party security assessments as a condition of catalog inclusion. This requirement creates a governance-aligned incentive structure: AI tools that have invested in demonstrable security posture can obtain fast-track approval, while tools that cannot demonstrate equivalent security must undergo additional review. It also provides an auditable record of the due diligence that shadow AI deployments structurally lack.

7.4 Zero Trust Guidance and AI Agent Identity

CSA's Zero Trust guidance provides the architectural foundation for applying continuous verification to AI agent identity and access. The classical Zero Trust principle of "never trust, always verify" translates directly to AI agent governance: agent credentials should be validated at every access request, not implicitly trusted based on the identity of the deploying employee; access scope should be enforced at the data and API layer, not merely stated in policy; and behavioral monitoring should continuously verify that agent actions remain within their authorized scope.

CSA's AI Organizational Responsibilities framework addresses the governance structure required to maintain these controls over time — defining which roles are accountable for AI system oversight, how responsibility is distributed across procurement, security, legal, and business unit functions, and what escalation paths exist when shadow AI is discovered or when authorized AI systems exhibit unexpected behavior.

8. Conclusions and Recommendations

8.1 The Structural Challenge

Shadow AI infrastructure represents the enterprise's fastest-growing ungoverned attack surface, combining the data exposure risk of shadow IT with the autonomous action risk, persistent memory, and supply chain opacity of modern AI systems. The governance deficit is structural: organizations lack the discovery tools, control frameworks, and policy architectures needed to manage AI adoption at the rate it is occurring, and employees have powerful productivity incentives to adopt AI capabilities through whatever channel is available — authorized or not.

The regulatory environment is closing the window for organizations to address this challenge on their own timeline. The EU AI Act's full applicability in August 2026 effectively mandates AI inventory and governance as a compliance requirement rather than a security best practice. NIST AI RMF adoption is accelerating across regulated industries, and sector-specific regulators in financial services, healthcare, and critical infrastructure are increasingly incorporating AI governance requirements into examination and compliance frameworks.

8.2 Prioritized Recommendations

Addressing shadow AI infrastructure requires action across three sequential horizons, each building on the foundation established by the previous. The immediate priority is gaining visibility — enterprises cannot govern what they cannot see, and the discovery gap is the single most consequential enabler of shadow AI risk. Near-term efforts translate that visibility into policy and classification infrastructure. Strategic actions embed governance permanently into organizational process so that new AI deployments are governed by design rather than caught retrospectively.

Immediate Actions (0–90 days): Deploy automated AI discovery tooling capable of enumerating shadow AI deployments across SaaS environments, identity provider OAuth grants, and network egress patterns. Establish an emergency triage process for high-risk discoveries — shadow AI deployments with broad data access, autonomous action authority, or processing of regulated data categories. Revoke OAuth grants for unauthorized AI applications pending review. Implement DLP policies covering AI assistant inputs and outputs as a technical enforcement backstop.

Near-Term Actions (90 days–12 months): Establish an approved AI tool catalog with clear minimum security requirements, a lightweight intake and review process, and transparent criteria for inclusion. Develop a risk classification framework for AI systems based on data access scope, action authority,

persistence, and supply chain transparency. Implement regular OAuth grant audits tied to the identity provider lifecycle management process. Train security operations staff to recognize AI-specific attack signatures, particularly prompt injection indicators in log data.

Strategic Actions (12+ months): Embed AI governance into the software development lifecycle, procurement process, and change management so that new AI capabilities are governed by design. Integrate AICM control domains into security assessment and vendor due diligence processes. Adopt CSA MAESTRO for threat modeling new agentic AI deployments. Pursue STAR for AI attestations or equivalent third-party assessments for high-priority AI tools. Establish behavioral monitoring capability for authorized AI agents that can detect out-of-scope actions and anomalous data access patterns.

8.3 The Governance Orientation

The most important strategic choice enterprises face is whether to approach AI governance as a prohibition regime or as a risk management regime. Prohibition has a strong intuitive appeal — eliminating unauthorized AI use eliminates the risks that use creates — but evidence consistently shows it does not work and imposes real competitive costs. Risk management accepts that employees will use AI, designs governance systems that channel that use toward lower-risk tools and behaviors, and applies technical controls that enforce minimum safety requirements regardless of which tools employees choose.

The organizations that will navigate the shadow AI governance challenge most effectively will not be those that ban AI most aggressively. They will be those that build the fastest, clearest pathways from "employee wants to use an AI tool" to "AI tool in use under governance," reducing the friction between demand and compliance until the incentive to bypass governance disappears.

References

[1] Reco AI. "2025 State of Shadow AI Report." Reco AI Research, 2025.

[2] SecurityWeek. "The Shadow AI Surge: Study Finds 50% of Workers Use Unapproved AI Tools." SecurityWeek, 2025.

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

[4] IBM. "2025 Cost of a Data Breach Report." IBM Think, 2025.

[5] Polymer. "Rogue AI Agents: What They Are and How to Stop Them." Polymer Security Blog, 2025.

[6] VentureBeat. "The Enforcement Gap: 88% of Enterprises Reported AI Agent Security Incidents." VentureBeat, 2025.

[7] Thehackernews. "The Hidden Security Risks of Shadow AI in Enterprises." The Hacker News, April 2026.

[8] OWASP. "LLM01:2025 Prompt Injection – OWASP Gen AI Security Project." OWASP, 2025.

[9] Proofpoint. "Cybersecurity Stop of the Month: How Threat Actors Weaponize AI Assistants with Indirect Prompt Injection." Proofpoint, 2025.

[10] Reco AI. "The Shadow AI Data Leak Problem." Reco AI Blog, 2025.

[11] Venn. "AI Data Leakage: What It Is and How to Protect Your Organization." Venn, 2025.

[12] Qualys. "Data Leakage Prevention in AI." Qualys Blog, April 2025.

[13] FireTail. "AI Incident Tracker: Breaches and Vulnerabilities." FireTail, 2025.

[14] OWASP. "OWASP Top 10 for Agentic Applications 2026." OWASP, 2026.

[15] IBM Newsroom. "IBM Report: 13% of Organizations Reported Breaches of AI Models or Applications, 97% of Which Reported Lacking Proper AI Access Controls." IBM, 2025.

[16] Thehackernews. "Shadow AI Discovery: A Critical Part of Enterprise AI Governance." The Hacker News, September 2025.

[17] Reco AI. "AI and Cloud Security Breaches: 2025 Year in Review." Reco AI Research, 2025.

[18] Sentra. "EU AI Act Compliance: What Enterprise AI Deployers Must Know." Sentra, 2025.

[19] NIST. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." NIST AI 100-1, January 2023.

[20] Credo AI. "Credo AI Named Gartner Cool Vendor 2025 in AI Cybersecurity Governance." Credo AI Blog, 2025.

[21] Reco AI. "What Is AI Sprawl and Why Is It a Growing SaaS Security Risk." Reco AI Learn, 2025.

[22] Microsoft. "Reduce Autonomous Agentic AI Risk." Microsoft Learn, 2025.

[23] Cloud Security Alliance. "AI Gone Wild: Why Shadow AI Is Your IT Team's Worst Nightmare." CSA Blog, March 2025.

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

[25] Morphisec. "Shadow AI: The Fastest-Growing Security Risk No One Is Tracking." Morphisec Blog, 2025.

[26] ISACA. "The Rise of Shadow AI: Auditing Unauthorized AI Tools in the Enterprise." ISACA, 2025.

[27] Securiti. "What Is Shadow AI? Risks, Examples, and Governance." Securiti, 2025.

[28] Palo Alto Networks Unit 42. "When AI Agents Go Rogue: Agent Session Smuggling Attack in A2A Systems." Unit 42 Blog, 2025.

[29] E-Security Planet. "77% of Employees Leak Data via ChatGPT." E-Security Planet, 2025.

[30] Lakera. "Prompt Injection and the Rise of Prompt Attacks." Lakera Blog, 2025.