CSAI Foundation

Cloud Security Alliance AI Safety Initiative

The Invisible Enterprise: Shadow AI and the Ungoverned Frontier

Asset Blindness, Systemic Risk, and the Imperative for AI Governance

Unofficial AI-assisted Research

2026-04-02

Table of Contents

Executive Summary

Enterprise AI adoption has dramatically outpaced enterprise AI governance. Across industries, employees are embedding AI assistants into daily workflows, developers are pulling open-source models into production pipelines without formal review, and business units are licensing AI SaaS platforms well outside the procurement visibility of their information security teams. The result is an invisible layer of AI infrastructure — broadly called Shadow AI — that operates without inventory, policy enforcement, access controls, or audit capability.

The scale of the problem is substantial. According to Reco's 2025 State of Shadow AI Report, 91% of AI tools in enterprise environments operate outside IT control, and organizations average 269 shadow AI applications per 1,000 employees [1]. IBM's 2025 Cost of a Data Breach Report found that one in five organizations experienced a breach involving shadow AI, and that such incidents cost an average of $670,000 more than the global breach average [2]. Menlo Security's 2025 enterprise analysis documented a 68% surge in shadow generative AI usage in a single year, with 57% of shadow AI users entering sensitive company data into unauthorized platforms [3].

The regulatory environment is narrowing. The EU AI Act's full compliance obligations take effect in August 2026, with mandatory AI system inventories as a prerequisite for any risk classification or conformity assessment [4]. ISO/IEC 42001 and the NIST AI Risk Management Framework similarly require organizations to govern AI assets as a foundational capability before more advanced controls can be applied [5][6]. Organizations that have not yet established comprehensive AI asset discovery and governance programs face mounting compliance exposure on top of ongoing security risk.

This whitepaper examines the anatomy of shadow AI, the specific security risks that emerge when AI assets go ungoverned, the real-world incidents that have materialized from this gap, and a practical governance architecture that security leaders can deploy to bring their AI estate under meaningful control. The recommendations throughout are grounded in CSA's own framework ecosystem, including the AI Controls Matrix (AICM), MAESTRO threat modeling methodology, the Capabilities-Based Risk Assessment (CBRA) framework, and CSA's AI Organizational Responsibilities series.

Introduction: The Proliferation Problem

The enterprise AI landscape of 2026 bears little resemblance to the AI landscape of even three years ago. What was once a capability available only through expensive, purpose-built enterprise contracts has become a consumer-grade commodity accessible to any employee with a web browser and a credit card.

ChatGPT, Gemini, Claude, Copilot, Perplexity, and hundreds of specialized vertical AI tools are used daily for tasks ranging from summarizing legal contracts to writing code, drafting communications, analyzing financial models, and interpreting customer data. Each of these tools represents a potential conduit for sensitive enterprise information to flow outward, and each represents a potential integration point for adversarial content to flow inward.

The phenomenon is not limited to consumer-facing AI assistants. Developers across organizations routinely download open-source models from repositories like Hugging Face and integrate them into production systems without formal security review. Engineering teams connect agentic frameworks to corporate data stores, messaging platforms, and API endpoints through Model Context Protocol (MCP) servers configured in hours, not weeks. Data science teams fine-tune models on proprietary datasets without engaging privacy or compliance review. The pace of AI adoption within the enterprise has outrun the governance mechanisms most organizations built for cloud, mobile, and SaaS adoption in prior eras — and the consequences are beginning to materialize.

Security teams are operating with a fundamental disadvantage: they cannot defend assets they do not know exist. When a database server goes unpatched, it is because asset management failed to register it. When an AI model processes customer PII without data classification controls, it is because no governance process was ever applied to it. The challenge of shadow AI is, at its core, an asset management and visibility problem — but one that carries risks unique to AI systems, including training data exfiltration, inference-time data leakage, supply chain compromise through malicious model weights, and adversarial manipulation of ungoverned inference pipelines.

This whitepaper uses the term "shadow AI" to describe any AI system, model, tool, API, agent, or integration that operates within an organization's environment without formal registration, risk assessment, policy governance, or security monitoring by an authorized team. The definition is intentionally broad: it encompasses the employee using ChatGPT to process customer emails, the developer who fine-tuned a Llama derivative on proprietary code without a security review, the business analyst who wired a third-party AI summarization API directly into Salesforce, and the operations team running an agentic workflow that autonomously queries corporate databases. All of these represent ungoverned AI assets, and all carry risk.

Defining Shadow AI: Anatomy of an Invisible Problem

Shadow AI is not a monolithic phenomenon. It manifests across multiple organizational layers and technical surfaces, each with distinct risk characteristics and governance requirements. Understanding the topology of shadow AI is a prerequisite for designing controls that address the full scope of the problem rather than its most visible symptoms.

The most familiar manifestation is consumer-facing AI assistant use — employees accessing platforms like ChatGPT, Gemini, or Claude through personal or corporate browsers for work tasks. This is widespread. Reco's enterprise telemetry across thousands of organizations found that 71% of office workers use AI tools without IT approval [1]. The Cybersecurity Dive reporting on UpGuard's survey found that 60.2% of white-collar employees have used AI tools at work, yet only 18.5% are aware of any official company policy regarding AI use [7]. Perhaps most striking is the executive dimension: 69% of presidents and C-suite executives and 66% of directors and senior VPs reported prioritizing speed over security in their AI tool adoption, suggesting that senior leaders may be among the most likely to bypass governance in favor of speed when adopting AI tools [7].

A second, less visible category comprises AI-native SaaS applications procured at the business unit level without IT involvement. These are not typically consumer tools but purpose-built platforms — AI-powered CRMs, legal review tools, HR screening systems, financial analysis assistants — licensed by individual departments who treat AI capability as a feature rather than a separate governance consideration. The AI processing within these platforms may handle sensitive employee data, customer records, or confidential business information, yet the underlying AI systems may never have been assessed for data residency, model training practices, or third-party subprocessor arrangements.

A third category is developer-sourced AI infrastructure: models downloaded from public repositories, AI SDKs embedded in application code, MCP servers deployed without change management review, and inference API connections established in development environments that subsequently migrate to production. JFrog's 2025 Software Supply Chain Report documented a 6.5-fold increase in malicious models on Hugging Face in a single year, with over one million new models added to that repository in 2024 alone [8]. An organization whose developers freely pull models from public repositories without verification is exposed to this supply chain risk in direct proportion to its lack of governance.

The fourth and most complex category is agentic AI — autonomous systems that not only generate text but take actions in connected systems, browse the web, execute code, query databases, send communications, and chain multiple AI capabilities together in pursuit of an objective. Agentic AI introduces risks that extend well beyond data leakage: an ungoverned agent may acquire permissions far in excess of its stated function, create persistent access mechanisms, or be hijacked through adversarial instructions embedded in documents or external content it processes. CSA's Agentic AI Red Teaming Guide identifies 12 critical vulnerability categories specific to autonomous AI agents, including authorization hijacking, permission escalation, hallucination exploitation, and chained access vulnerabilities [9]. Ungoverned agentic deployments are particularly dangerous because the attack surface extends to every external system the agent can reach, and the agent's behavior may be invisible to both the user who deployed it and the security team that was never informed.

Asset Blindness: The AI Inventory Crisis

The most fundamental governance failure underlying shadow AI is the absence of comprehensive AI asset inventory. Organizations cannot classify, control, or audit AI systems they do not know exist. Yet asset discovery for AI is significantly harder than for traditional IT assets, because AI systems leave faint and inconsistent signals in the infrastructure telemetry that security operations centers were built to monitor.

A virtual machine registers in a cloud console. A SaaS application appears in an SSO log. An AI tool accessed through a personal browser, invoked via a direct API key embedded in application code, or installed as a browser extension may leave no trace in any of these systems. Reco's telemetry data found that some shadow AI tools had median usage durations of over 400 days without formal approval — not brief experiments but deeply embedded operational dependencies that had grown invisible roots in organizational workflows [1]. When such tools are eventually discovered, removing them creates operational disruption that creates political pressure to continue tolerating the ungoverned deployment.

The scale of organizational blindness is documented across multiple independent studies. A 2025 survey of 461 security professionals found that 83% of organizations lack automated AI controls, leaving sensitive data exposed to public AI tools without visibility [10]. Only 42% of organizations fully understand the types of AI in use within their current security stack [10]. IBM's 2025 breach data found that only 37% of organizations have policies to manage AI or detect shadow AI, and that of those, only 34% conduct regular audits for unauthorized AI use [2]. In aggregate, these figures describe an enterprise sector in which the overwhelming majority of organizations are making security decisions about an AI landscape they cannot see.

The challenge is compounded by the heterogeneity of what must be inventoried. A complete AI asset inventory for a mid-sized enterprise in 2026 might need to account for AI-capable API endpoints integrated into business applications, large language models running on internal or cloud-hosted GPU infrastructure, fine-tuned model derivatives derived from foundation models but modified with proprietary data, browser extensions and productivity plugins with embedded AI capabilities, third-party vendors whose services include AI-powered processing of shared data, agentic frameworks with connections to internal data stores and external services, and MCP servers connecting AI systems to corporate tools. Each asset class has distinct discovery mechanisms, risk profiles, and governance requirements. None of them are well-served by traditional asset management tools designed for IP-addressed endpoints and installed software inventories.

This technical heterogeneity creates what might be called a governance gap stack: even organizations that have recognized the shadow AI problem and begun to address it systematically face years of discovery, classification, and remediation work. The gap between organizational AI deployment velocity and organizational AI governance maturity appears to be widening rather than narrowing, precisely because AI capability is becoming easier to deploy while AI governance capability remains complex and resource-intensive to build.

The Attack Surface of Ungoverned AI

When AI assets operate outside governance frameworks, they do not merely create compliance exposure — they actively expand the organizational attack surface in ways that are qualitatively different from traditional ungoverned IT assets. Four categories of risk deserve particular attention.

Data Exfiltration Through AI Interfaces

The most immediately material risk of shadow AI is the exfiltration of sensitive data through AI interfaces that employees do not perceive as data-sharing acts. When an engineer pastes proprietary source code into ChatGPT to debug it, or a lawyer uploads a draft merger agreement to an AI review tool, or a sales representative enters a client's financial details into a customer analysis assistant, sensitive organizational data is transmitted to third-party systems whose data handling, training practices, and security posture may be entirely unknown to the organization.

Samsung's experience in early 2023 provided the canonical illustration of this pattern. Without any formal authorization, employees in the semiconductor division had adopted ChatGPT for productivity tasks; within weeks, three separate data exfiltration incidents occurred: an engineer pasted faulty source code for debugging; another entered production equipment code to request optimizations; and a third converted a recorded internal meeting into text for summarization [11]. Samsung banned generative AI tools company-wide in May 2023 after discovering these incidents. The Samsung case is notable precisely because it was not the result of malicious intent — it was ordinary, productivity-motivated behavior that happened to route confidential intellectual property to an external AI system.

Cisco's 2025 Data Privacy Benchmark Study found that 46% of organizations reported internal data leaks through generative AI applications, including employee names, internal data, and business-sensitive information [12]. Menlo Security's 2025 analysis documented 155,005 copy attempts and 313,120 paste attempts to AI platforms in a single monitored month across its enterprise customer base, with source code accounting for 42% of policy violations and regulated data accounting for 32% [3]. The volume of AI-directed data flows that Menlo Security documented in its 2025 enterprise analysis significantly exceeds the review capacity of any manual monitoring program, reinforcing the case for automated classification and enforcement.

Vulnerability Exploitation in AI-Integrated Systems

The integration of AI capabilities into enterprise workflows creates new attack vectors that adversaries are actively exploiting. A June 2025 vulnerability in Microsoft 365 Copilot allowed attackers to embed malicious instructions within received emails, enabling zero-click exfiltration of sensitive corporate data without any user action [13]. A separate vulnerability in Microsoft Copilot Studio, disclosed by Lasso Security, exposed

sensitive cached repository data from an estimated 16,290 organizations through improperly managed historical content access [14]. These are not hypothetical risks — they are documented compromises affecting production enterprise AI deployments at scale.

Ungoverned AI systems are particularly vulnerable to prompt injection attacks, in which adversarial instructions embedded in content the AI processes cause the system to take actions its deployers did not intend or authorize. An agentic AI configured to summarize incoming emails may be instructed via a malicious email to forward all subsequent correspondence to an external address. An AI assistant integrated with a corporate document management system may be induced to retrieve and transmit confidential files through instructions embedded in a document a user asks it to analyze. These attacks are effective precisely because the AI system has access to organizational resources — access that was granted informally, without the policy controls that would otherwise govern a human employee with the same data access.

Supply Chain Compromise Through Model Repositories

Open-source AI model repositories have become a significant supply chain attack vector. As documented in the prior section, the JFrog 2025 Software Supply Chain Report found a 6.5-fold year-over-year increase in malicious models on Hugging Face, spanning the full spectrum from models containing serialized Python payloads that execute on loading to models with embedded backdoors activated under specific input conditions [8]. The threat actor group NullBulge weaponized repositories on both Hugging Face and GitHub using Python payloads that exfiltrated data via Discord webhooks and delivered LockBit ransomware variants to downstream users [29].

Palo Alto Networks Unit 42 documented the "model namespace reuse" attack pattern, in which malicious actors re-register abandoned model namespaces on platforms like Hugging Face, causing production pipelines that fetch models by reference name to silently receive attacker-controlled versions [16]. Organizations whose development teams pull models from public repositories without formal verification are exposed to this risk in direct proportion to how frequently they update or experiment with new models. In environments where developers have freedom to integrate new AI capabilities without a security review process, the path from public repository to production inference pipeline can be a matter of hours.

Agentic AI: Autonomous Risk Accumulation

Agentic AI systems introduce a qualitatively distinct risk category because their risk profile is not static — it grows as the agent acquires access, builds history, and embeds itself in operational workflows. An agentic system deployed without governance may gradually accumulate permissions to organizational resources far beyond what its original task required, because granting it new access is the path of least resistance for the employees who use it. CSA's CBRA framework specifically identifies "autonomy" and "permission scope" as

multiplicative risk factors: an AI system that operates autonomously and has broad permissions to organizational resources presents significantly higher risk than one that is constrained in either dimension [17].

The "shadow agent" problem — agentic AI systems deployed without security team knowledge — represents the frontier of the shadow AI challenge. Unlike a chat assistant that a user invokes deliberately, an agent may run continuously, take actions across multiple systems, generate persistent artifacts, and accumulate a history of operations that would be difficult to audit even if security teams were aware of its existence. In the absence of governance, the agent's behavior is governed only by its instructions and the capabilities of its model — neither of which provides the accountability structures that organizational security requires.

Real-World Impact: The Cost of Unmanaged AI

Risk management frameworks define risk as the product of likelihood and impact — and for shadow AI, both dimensions are elevated. IBM's 2025 Cost of a Data Breach Report provides the most comprehensive financial accounting of shadow AI's organizational cost: breaches involving shadow AI added an average of $670,000 to incident costs compared to the global average, and shadow AI incidents were disproportionately likely to compromise personally identifiable information (65% of shadow AI breaches versus a global average of 53%) and intellectual property (40% versus a global average of 33%) [2]. The report found that 97% of organizations breached through AI systems lacked proper AI access controls — suggesting that the presence of basic governance may have been protective in many of these cases, or at minimum would have reduced the exposure window [2].

The detection window for shadow AI incidents is also longer than for conventional security events. IBM's data indicates that shadow AI breaches had a mean time to identify and contain of 247 days, compared to an already-concerning 241-day average across all breach types [2]. This extended dwell time is consistent with the structural challenge: shadow AI systems are often not monitored at all, meaning that anomalous behavior in those systems generates no alerts. In many cases, such breaches surface through downstream effects — a regulator's inquiry, a customer complaint, a threat intelligence report — rather than through proactive security detection, consistent with the broader pattern of extended dwell times documented in breach data [2].

Gartner's analysis of trajectory is equally sobering. The firm projects that by 2030, more than 40% of enterprises will experience a security or compliance incident linked to unauthorized shadow AI [18]. A separate Gartner prediction from early 2025 holds that 40% of AI data breaches will arise from cross-border generative AI misuse by 2027 [19] — a figure directly relevant to organizations with multinational operations that are deploying AI tools without attention to the data residency and cross-border transfer implications that multiple regulatory regimes now require them to address.

The financial impact extends beyond direct breach costs. Organizations found to be operating high-risk AI systems without the required risk assessments, documentation, and controls under the EU AI Act face fines of up to €30 million or 6% of global annual turnover, whichever is higher, for violations involving prohibited AI practices [4]. For organizations that have not yet inventoried their AI estate, the risk of inadvertently operating a system that qualifies as high-risk under the Act's definitions — without the required conformity assessment — is real and growing as the August 2026 compliance deadline approaches.

The Regulatory Mandate: Compliance Windows Are Closing

The regulatory environment around AI governance has moved from aspirational guidance to binding obligation with enforceable penalties. Organizations that continue to treat AI governance as a future investment rather than a present operational requirement are accumulating regulatory risk that compounds with each passing month.

The EU AI Act entered into force on August 1, 2024, establishing the world's first comprehensive horizontal AI regulatory framework with hard legal obligations [4]. The compliance timeline is phased but accelerating: prohibitions on banned AI practices took effect February 2, 2025; obligations for general-purpose AI models became applicable August 2, 2025; and full obligations for high-risk AI systems, including conformity assessments, technical documentation, and post-market monitoring, take effect August 2, 2026. For organizations that have not yet completed an AI inventory and risk classification exercise, the timeline to August 2026 is compressed. The Act's Article 51 requires that AI systems meeting high-risk criteria be registered in the EU database before they are placed on the market or put into service — a requirement that is impossible to fulfill for AI systems the organization does not know it is operating.

The NIST AI Risk Management Framework (AI RMF), while voluntary in the United States, has been cited as a reference framework in sector-specific guidance from financial regulators, and is widely referenced in risk management best practice documentation across regulated industries [5]. The AI RMF's four core functions — Govern, Map, Measure, and Manage — establish a logical dependency: an organization cannot measure or manage the risk of AI systems it has not mapped, and it cannot map systems it has not inventoried. ISO/IEC 42001, the certifiable international standard for AI management systems, similarly treats AI inventory and documentation as foundational requirements that precede all other management system activities [6].

The table below summarizes the key regulatory instruments and their primary AI inventory and governance requirements.

Regulation / Standard	Jurisdiction	AI Inventory Requirement	Primary Compliance Deadline
EU AI Act	European Union	Mandatory inventory and risk classification of all AI systems; high-risk AI registration in EU database	August 2, 2026 (high-risk obligations)
NIST AI RMF	United States (voluntary)	MAP function: identify and categorize all AI systems and their context of use	Ongoing; incorporated by sector regulators
ISO/IEC 42001	International	Documented AI system inventory as prerequisite to AI management system certification	Ongoing; certifiable against audited criteria
NY DFS Circular (AI Guidance)	New York State	AI risk assessments and governance documentation for regulated financial entities	Effective 2024 [30]
EU DORA (ICT Risk for Finance)	European Union	ICT asset management including AI systems embedded in financial processes	Effective January 17, 2025 [31]

The convergence of these requirements across jurisdictions means that multinational organizations face an overlapping matrix of obligations that all share a common prerequisite: knowing what AI systems they operate. This prerequisite is not currently met in the majority of enterprises, as the data cited throughout this paper demonstrates.

Building Visibility: AI Asset Discovery and Inventory

The path from shadow AI blindness to managed AI governance begins with discovery. Organizations must approach AI asset inventory as an ongoing operational capability, not a one-time project, because the AI landscape within any enterprise changes continuously as employees adopt new tools, developers integrate new models, and vendors embed new AI capabilities in existing platforms.

Effective AI asset discovery requires operating across multiple discovery channels simultaneously. Network traffic analysis can surface connections to known AI service endpoints — API calls to OpenAI, Anthropic, Hugging Face inference endpoints, and major cloud AI services that are not registered in the organization's approved vendor list. SaaS discovery tools that connect to identity providers and monitor OAuth authorizations can surface AI applications that employees have granted access to organizational accounts. Code repository scanning can identify AI SDK imports, model download scripts, and embedded API keys that indicate AI integrations in development or production code. Cloud infrastructure inventory tools adapted for AI workloads — such as Qualys TotalAI, Wiz's AI security module, and Pillar Security's asset inventory capabilities — can surface GPU-backed compute resources, inference endpoints, and container-based model deployments in cloud environments [20][21][22].

Each discovery channel captures a different slice of the shadow AI landscape, and none of them is comprehensive alone. An integrated AI asset management capability must aggregate findings across all channels into a unified inventory that tracks each asset's identity, function, data access, owner, deployment context, and governance status. Critically, the inventory must be maintained as a living record: as noted in the asset blindness analysis above, shadow AI tools routinely embedded themselves into workflows over 400-day periods without detection [1] — a pattern that reflects what happens when discovery is treated as an event rather than a process.

Once discovered, assets must be classified according to their risk profile. CSA's Capabilities-Based Risk Assessment framework provides a structured scoring methodology that evaluates each AI system across four dimensions: Criticality (how important the system's function is to organizational operations), Autonomy (the degree to which the system acts independently of human oversight), Permission Scope (the breadth of organizational resources the system can access), and Potential Impact (the severity of harm the system could cause if compromised or misused) [17]. The multiplicative scoring model produces a risk tier — Low, Medium, or High — that directly drives the level of AICM controls applied to each system. This approach ensures that governance investment is proportional to risk, avoiding both over-engineering controls for low-risk assistive tools and under-engineering them for high-stakes autonomous systems.

The organizational dimension of asset discovery is as important as the technical one. Many AI deployments begin with informal business decisions that never touch security or IT processes. Effective shadow AI governance programs therefore combine technical discovery with organizational outreach: published AI use policies with clear registration requirements, self-service intake processes that make registering an AI tool easier than routing around governance, and periodic surveys of business units to surface AI tool usage that technical discovery cannot reach. Evidence from shadow IT governance programs suggests that organizations that position governance as a service — helping employees use AI tools safely rather than simply policing their use — achieve meaningfully higher voluntary disclosure rates than those that approach the problem primarily through enforcement.

Governance Architecture: From Discovery to Control

Asset discovery and inventory are the foundation of AI governance, but they are not sufficient on their own. The governance architecture that sits on top of the inventory must address policy, access control, monitoring, incident response, and continuous improvement. CSA's AI Controls Matrix provides a structured reference for this architecture across 18 security domains and more than 240 control objectives [23].

Policy and Organizational Structure

Effective AI governance requires designated accountability. Organizations should establish a formal AI Risk Review Board with representation from security, legal, compliance, privacy, and relevant business units, charged with reviewing and approving all new AI system deployments above a threshold risk score. The board's charter should specify the intake process for new AI tools, the documentation requirements for registration, the review cadence for existing AI assets, and the escalation path for governance exceptions. This structure mirrors the established model for cloud service procurement governance and can often be built on existing vendor risk management processes adapted for AI-specific considerations.

Acceptable Use Policies for AI must be specific enough to provide practical guidance while flexible enough to accommodate legitimate business needs. In practice, policies that simply prohibit "unauthorized AI use" without specifying what authorization requires, what categories of data may not be processed by AI tools, and what constitutes a reportable incident tend to be widely ignored, because they provide employees with no actionable path to compliant behavior. Effective policies enumerate the categories of data that may not be entered into external AI systems (customer PII, protected health information, material non-public financial data, source code subject to license restrictions, merger-related information), specify the approved channels through which AI tools may be adopted, and establish clear consequences for policy violation alongside clear amnesty provisions for employees who self-disclose existing shadow AI use.

Access Controls and Zero Trust Architecture

The access control posture for AI systems should be governed by Zero Trust principles: no AI system should be granted access to organizational resources on the basis of assumed trust, and all access should be continuously validated against current policy. In practice, this means that AI systems should be provisioned with minimum-necessary data access, with access scoped to specific datasets, time windows, and functional purposes rather than granted broadly on the basis of an AI system's overall registration status. CSA's work on Shadow Access — specifically the intersection of AI deployments and identity security — identifies the phenomenon of AI systems that accumulate permissions over time through informal expansion as one of the primary drivers of AI-related access risk [24].

Service accounts and API keys provisioned for AI systems represent a particularly high-value target for credential theft and a particularly high-risk category of credential sprawl. Each AI system in the inventory should be associated with a set of explicit, minimal-privilege credentials, and those credentials should be subject to the same lifecycle management, rotation policies, and monitoring as human identity credentials. The credential inventory for AI systems should be a component of the broader AI asset registry, enabling security teams to detect when AI systems are using credentials beyond their intended scope or when credentials associated with deprovisioned AI systems remain active.

Continuous Monitoring and Behavioral Baselines

Monitoring for AI systems must address both the systems themselves and the data flows they generate. At the system level, organizations should establish behavioral baselines for registered AI systems — the normal patterns of API call volume, data access, output generation, and integration activity — and alert on deviations that could indicate compromise, misuse, or unexpected capability expansion. CSA's MAESTRO threat modeling framework provides a structured approach to identifying the specific behaviors that constitute threat indicators for different categories of AI system, organized across the seven-layer MAESTRO model spanning model foundations through ecosystem integration [25].

At the data flow level, Data Loss Prevention capabilities must be extended to cover AI interfaces. Many traditional DLP solutions were designed for email, web uploads, and endpoint file transfers, and require adaptation to monitor the data flows between organizational users and AI API endpoints, to classify the content of prompts sent to AI systems, and to detect the transmission of sensitive information types that organizational policy prohibits from AI processing. The 313,120 paste operations to AI platforms in a single monitored month that Menlo Security documented in its 2025 enterprise analysis [3] illustrates why manual review of AI-directed data flows is not feasible — automated classification and enforcement are required.

Supply Chain Security for AI Models

Organizations that develop or fine-tune AI models, or that deploy open-source models in production systems, require an AI-specific supply chain security program that complements their existing software supply chain controls. This program should mandate cryptographic verification of model provenance before deployment, using checksums or signed manifests to verify that a downloaded model matches the version published by a trusted source and has not been modified in transit or at rest. AI Software Bills of Materials (AI-SBOMs) — structured records of a model's training data provenance, architecture, fine-tuning history, and dependency chain — should be required for all production AI deployments and should be incorporated into vendor due diligence assessments for third-party AI services.

The model namespace reuse attack pattern identified by Palo Alto Networks Unit 42 [16] and the malicious model upload campaigns documented by Checkmarx [15] demonstrate that supply chain attacks against AI models are not theoretical but active and evolving. Organizations should treat model repositories as untrusted sources requiring the same verification and sandboxed evaluation process applied to third-party code libraries, and should maintain an approved model registry — an internal catalog of verified, approved model versions — that development pipelines are required to use rather than fetching models directly from public repositories.

CSA Resource Alignment

CSA has developed an integrated ecosystem of frameworks and guidance that collectively address the shadow AI and AI governance challenge. Organizations building governance programs should map their activities directly to this ecosystem to ensure alignment with emerging regulatory expectations and industry best practice.

The **AI Controls Matrix (AICM)** provides the most comprehensive control reference for enterprise AI governance, spanning 18 domains from AI model security and data governance through supply chain risk and incident response [23]. The AICM's 240+ control objectives can be applied proportionally based on the risk tier assigned through CBRA scoring, ensuring that organizations invest governance effort in proportion to actual risk. The AICM is aligned with the EU AI Act, NIST AI RMF, and ISO/IEC 42001, helping organizations use a unified control reference to address the shared control objectives across multiple regulatory regimes, reducing redundancy in compliance implementation.

The **Capabilities-Based Risk Assessment (CBRA) for AI Systems** framework provides the scoring methodology that determines which AICM control tiers apply to each AI system in the inventory [17]. CBRA's four-dimensional scoring model — Criticality, Autonomy, Permission Scope, and Potential Impact — is specifically designed to address the risk characteristics of AI systems that distinguish them from conventional IT assets, including the autonomy dimension that is particularly relevant for agentic deployments.

The **MAESTRO threat modeling framework** addresses the specific threat landscape of agentic and large language model-based AI systems, providing a structured approach to identifying adversarial scenarios across seven layers of the AI stack [25]. Security teams governing AI systems should use MAESTRO to develop threat models for high-risk AI deployments, informing both the monitoring strategy and the incident response playbooks for each system.

The **AI Organizational Responsibilities** series — covering AI Tools and Applications, Core Security Responsibilities, and Governance, Risk Management, Compliance and Cultural Aspects — provides implementation guidance for the organizational structures, roles, and processes that underpin technical AI governance [26][27][28]. The series explicitly addresses shadow AI prevention as an organizational governance objective and provides RACI templates for distributing AI security responsibilities across the functions that must collaborate to govern AI effectively.

The **Shadow Access and AI** publication and the companion **Confronting Shadow Access Risks** guidance address the specific intersection of AI deployments and identity security, providing a framework for managing the access accumulation and privilege escalation risks that ungoverned AI systems generate [24]. The Zero Trust architecture recommendations in these publications are directly applicable to the access control architecture for registered AI systems and to the detection of unauthorized AI access to organizational resources.

Conclusions and Recommendations

The shadow AI problem is not a future risk to be monitored — it is a present operational reality with documented financial, regulatory, and reputational consequences. Organizations that have not yet established comprehensive AI governance programs are already accumulating risk from AI systems they cannot see, and the regulatory environment is reducing the tolerance period for this posture to near zero.

The following recommendations are addressed to security leaders who need to establish or accelerate AI governance programs in 2026.

Establish a comprehensive AI asset inventory immediately. No governance capability is meaningful without knowing what assets it governs. Deploy technical discovery capabilities across network, identity, code repository, and cloud infrastructure channels to surface existing AI deployments, and establish a mandatory registration process for all new AI system deployments. Treat the inventory as an operational capability requiring continuous maintenance, not a one-time audit project.

Apply CBRA risk scoring to all discovered assets. Once the inventory has been populated, score each AI system on CSA's four-dimensional CBRA model to assign a risk tier and determine the appropriate level of AICM controls. Prioritize remediation effort toward high-risk ungoverned systems — particularly any agentic AI with broad data access and minimal oversight — while applying proportionally lighter governance to low-risk assistive tools that carry limited organizational exposure.

Extend Zero Trust architecture to AI systems explicitly. AI systems should be provisioned with minimum-necessary credentials, enrolled in continuous access validation, and subjected to the same behavioral monitoring as human users. Credentials associated with AI systems should be managed in the

organizational PAM system, rotated regularly, and immediately revoked when the associated AI system is deprovisioned. Access scope for each AI system should be documented in the asset registry and reviewed at every renewal cycle.

Implement AI-aware DLP and monitoring. Extend Data Loss Prevention capabilities to monitor data flows between organizational users and external AI service endpoints. Establish behavioral baselines for registered AI systems and alert on deviations. Integrate AI monitoring telemetry into the security operations center's threat detection workflow, with playbooks specific to AI-related incident patterns including data exfiltration through AI interfaces, prompt injection indicators, and anomalous model output patterns.

Establish supply chain security controls for AI models. Require cryptographic verification of model provenance before deployment, maintain an internal approved model registry, and mandate AI-SBOMs for all production AI deployments. Treat public model repositories as untrusted sources requiring sandboxed evaluation before integration into organizational pipelines. Extend vendor due diligence requirements to cover AI model providers and AI-enabled SaaS platforms, including assessment of their own AI supply chain security practices.

Accelerate EU AI Act compliance readiness. Organizations subject to EU AI Act obligations — which includes most multinational enterprises and any organization processing the data of EU residents — must complete AI system inventory and risk classification exercises before the August 2026 high-risk obligations deadline. Organizations that have not begun this process should treat it as a time-sensitive program, not a routine compliance project, given the complexity of inventorying existing AI deployments and the lead time required to implement conformity assessment processes for systems that qualify as high-risk.

Build governance as a service, not a gatekeeping function. The organizational culture dimension of shadow AI governance is as important as the technical dimension. Organizations that position AI governance as a friction-free service — providing employees and developers with fast, supported paths to adopt AI tools within sanctioned frameworks — will achieve higher voluntary compliance and earlier visibility into emerging AI use than organizations that rely primarily on enforcement. Establish a clear intake process, publish an approved AI tool catalog, offer AI security training as a positive enabler, and communicate governance requirements through channels that reach the employees who are deploying AI in practice.

The shadow AI problem is tractable. The organizations that will navigate it successfully are those that invest now in the visibility infrastructure — the inventory, monitoring, and governance architecture — that transforms an invisible AI estate into a managed one. The technology to do this exists. The frameworks to guide it exist. The regulatory imperative to act is clear and immediate. What remains is organizational commitment to treat AI governance as the core security capability that the current threat and regulatory environment demands.

References

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

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

[3] Menlo Security. "2025 Report: Shadow Generative AI Usage in the Modern Enterprise." Menlo Security, August 2025.

[4] European Commission. "Regulatory Framework for Artificial Intelligence." European Commission Digital Strategy, 2024.

[5] NIST. "AI Risk Management Framework (AI RMF 1.0)." National Institute of Standards and Technology, January 2023.

[6] ISO/IEC. "ISO/IEC 42001:2023 — Information Technology — Artificial Intelligence — Management System." International Organization for Standardization, 2023.

[7] Cybersecurity Dive. "Shadow AI: Employee Trust and Unauthorized Tool Use." Cybersecurity Dive, 2025.

[8] JFrog. "2025 Software Supply Chain State of the Union." JFrog, 2025.

[9] Cloud Security Alliance. "Agentic AI Red Teaming Guide." CSA AI Safety Initiative, 2025.

[10] Kiteworks. "AI Security Gap 2025: Organizations Flying Blind." Kiteworks, 2025.

[11] TechCrunch. "Samsung Bans Use of Generative AI Tools Like ChatGPT After April Internal Data Leak." TechCrunch, May 2023.

[12] Cisco. "2025 Data Privacy Benchmark Study." Cisco, April 2025.

[13] The Hacker News. "Zero-Click AI Vulnerability Exposes Microsoft 365 Copilot Data." The Hacker News, June 2025.

[14] Lasso Security. "Major Vulnerability in Microsoft Copilot Studio." Lasso Security Research, 2024.

[15] Checkmarx. "Hugs from Strangers: AI Model Confusion Supply Chain Attack." Checkmarx Security Research, 2025.

[16] Palo Alto Networks Unit 42. "Model Namespace Reuse: AI Supply Chain Attack Vector." Palo Alto Networks, 2025.

[17] Cloud Security Alliance. "Capabilities-Based Risk Assessment (CBRA) for AI Systems." CSA AI Safety Initiative, 2025.

[18] IT Pro. "Gartner: 40% of Enterprises Will Experience Shadow AI Breaches by 2030." IT Pro, November 2025.

[19] Gartner. "Gartner Predicts 40% of AI Data Breaches Will Arise from Cross-Border GenAI Misuse by 2027." Gartner Newsroom, February 2025.

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

[21] Wiz. "AI Inventory and AI Security in Cloud Environments." Wiz AI Security Academy, 2025.

[22] Pillar Security. "AI Asset Inventory: The Foundation of AI Governance and Security." Pillar Security Blog, 2025.

[23] Cloud Security Alliance. "AI Controls Matrix (AICM)." CSA AI Safety Initiative, 2024.

[24] Cloud Security Alliance. "Shadow Access and AI: Confronting Shadow Access Risks in Zero Trust and AI Deployments." CSA, 2025.

[25] Cloud Security Alliance. "MAESTRO: Multi-Agent Environment, Autonomy, and Risk for Threat Reasoning and Operations." CSA AI Safety Initiative, 2025.

[26] Cloud Security Alliance. "AI Organizational Responsibilities: AI Tools and Applications." CSA AI Safety Initiative, 2024.

[27] Cloud Security Alliance. "AI Organizational Responsibilities: Core Security Responsibilities." CSA AI Safety Initiative, 2024.

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

[29] SentinelOne. "NullBulge | Threat Actor Masquerades as Hacktivist Group Rebelling Against AI." SentinelOne Labs, 2024.

[30] New York Department of Financial Services. "Guidance on the Use of Artificial Intelligence and External Consumer Data and Information Sources." NYDFS, 2024.

[31] European Parliament and Council of the European Union. "Regulation (EU) 2022/2554 on Digital Operational Resilience for the Financial Sector (DORA)." Official Journal of the European Union, December 2022.