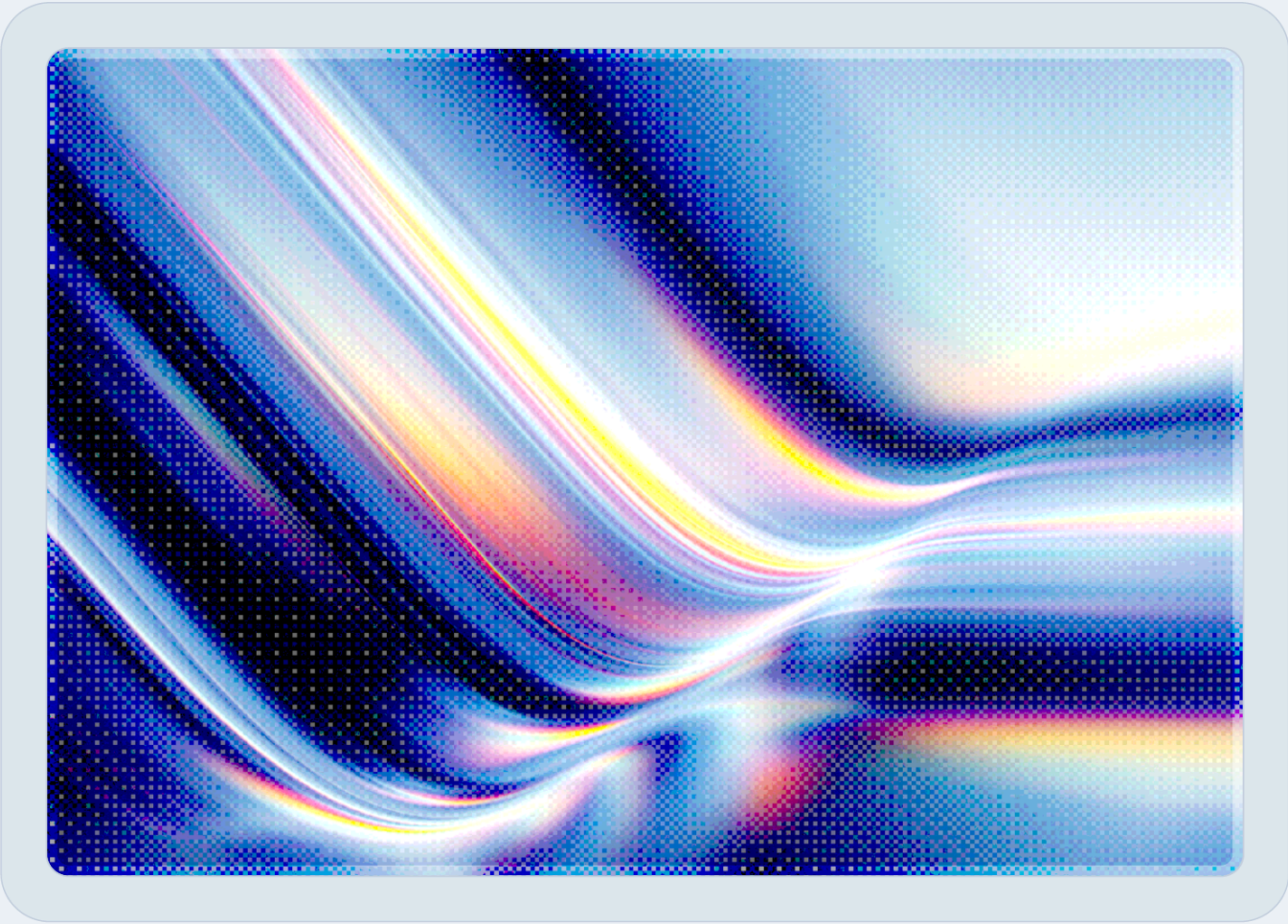
CSAI Foundation | Cloud Security Alliance

The AI Security Governance Gap

Shadow AI, Ownership Fragmentation, and the Enterprise Breach Blindspot

2026-05-20

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary

Enterprise generative AI adoption has, over the eighteen months between late 2024 and early 2026, outpaced governance by a wide and measurable margin. IBM's 2025 Cost of a Data Breach report, which draws on data from 600 organizations across 16 countries, found that one in five breached organizations now traces the incident to shadow AI — unauthorized employee use of AI tools outside the enterprise's sanctioned environment — and that the average shadow-AI-related breach costs $670,000 more than a breach where shadow AI was not a contributing factor [1][2]. The same report shows that thirteen percent of organizations reported a breach involving an AI model or AI application directly, and that ninety-seven percent of those organizations lacked the basic AI access controls that would have constrained the blast radius [1][3]. Sixty-three percent of breached organizations either had no AI governance policy in place or were still developing one [2].

We argue that these numbers reflect a structural condition rather than a transient adoption gap: the fragmentation of ownership and the absence of identity controls for AI principals are institutional failures that do not self-correct as tooling matures. The Cloud Security Alliance and Google Cloud's 2025 State of AI Security and Governance survey of three hundred IT and security professionals found that only twenty-six percent of organizations had comprehensive AI security governance policies, and that seventy-two percent of respondents were neutral or not confident in their organization's ability to execute an AI security strategy — down from fifty-one percent in 2024, when that share described themselves as confident or very confident [4]. The same survey identified fragmented AI ownership as a defining feature of the current state: while fifty-three percent of organizations say the security team is primarily responsible for securing AI, the responsibility for AI deployment is split across dedicated AI/ML teams (twenty percent), IT departments (nineteen percent), cross-functional groups (sixteen percent), security teams (thirteen percent), senior leadership (nine percent), and data science (eight percent) [4]. Funding authority for AI security tools is similarly distributed across CISO (forty-nine percent), CTO (thirty-six percent), CIO (thirty-three percent), business unit owners (twenty-five percent), and Chief AI Officers (eleven percent) [4]. The Boston Consulting Group's executive research has found that while eighty-five percent of executives agree AI is a top priority, only fourteen percent of organizations have clearly defined the leadership roles and responsibilities required to manage it [5]. The result is that AI risk has become a multi-owner concern with no single accountable owner.

The breach evidence accumulated through 2025 and into 2026 shows what happens when shadow adoption, fragmented ownership, and missing identity controls compound. EchoLeak, disclosed in June 2025 as CVE-2025-32711, demonstrated that an attacker could exfiltrate sensitive data from Microsoft 365 Copilot through a single crafted email with no user interaction required, chaining prompt injection past Microsoft's own classifier, past link redaction, and through an allow-listed Teams content-delivery path [6][7].

AgentFlayer, presented at Black Hat USA 2025 by Zenity Labs, demonstrated that the same class of zero-click compromise applied across OpenAI ChatGPT, Microsoft Copilot Studio, Salesforce Einstein, Google Gemini, and developer tools running with Jira MCP integrations [8]. ForcedLeak, disclosed in September 2025 as a 9.4-CVSS flaw in Salesforce Agentforce, allowed an attacker to exfiltrate CRM data through a malicious Web-to-Lead submission processed by a customer's AI agent [9]. The pattern across these incidents is not the sophistication of the underlying technique. It is the fact that the AI agent layer routinely operates with broad data access, weak input validation against attacker-controlled content, and identity privileges that no one has scoped to a discrete task.

This whitepaper analyzes the three reinforcing fault lines that produce what we call the AI Security Governance Gap. The first is shadow AI itself — the unmanaged channel through which employees move corporate data into consumer AI accounts that the enterprise cannot see, audit, or revoke. The second is ownership fragmentation — the diffusion of AI accountability across the CIO, CTO, CISO, Chief AI Officer, business units, legal, and compliance, with the practical consequence that no single role can be held to account when a model fails or a tool leaks. The third is the identity and access blindspot — the failure of existing identity, access, and shadow access controls to keep pace with the non-human principals that AI agents, MCP servers, and automated workflows have introduced into enterprise environments. The paper concludes with a control program aligned to the CSA AI Controls Matrix, MAESTRO threat-modeling framework, Confronting Shadow Access guidance, and AI Organizational Responsibilities publications, and with a set of practical recommendations that security leaders can implement in the current planning cycle to close the gap that the 2025 breach data now describes in numbers.

Introduction and Background

The phrase "governance gap" has been part of the AI security vocabulary since at least the early 2024 wave of generative AI adoption, but it has acquired specific empirical content over the last twelve months. Three forms of evidence now converge on the same conclusion. The first is breach data: organizations that lack AI governance experience measurably larger losses, and AI-specific attack patterns are now showing up in incident reports with sufficient frequency to enter the standard breach taxonomy [1]. The second is survey data: parallel research by CSA and Google Cloud, by Boston Consulting Group, by Grant Thornton, by IBM's X-Force, and by McKinsey has found, with substantial agreement, that the gap between AI ambition and AI governance maturity is — taken together across independent research bodies — the most acute of any technology adoption cycle in the cloud era, though no single study has conducted a direct historical comparison to prior waves [4][5][10][11]. The third is incident-level technical evidence: zero-click prompt injection, indirect prompt injection through trusted business records, and silent data exfiltration through AI agents have moved from research curiosities into production exploit chains affecting the major enterprise AI platforms [6][7][8][9].

The structural reason these patterns are emerging now, rather than during earlier waves of cloud and SaaS adoption, has three sources. First, generative AI lowered the cost of consuming AI from "build a model" to "open a tab" within eighteen months. By Q3 2025 LayerX, a browser-security vendor with commercial interest in AI-specific DLP and browser visibility tools, reported in its Enterprise AI and SaaS Data Security Report — drawing on browser telemetry from enterprise customer environments — that seventy-seven percent of employees paste data into generative AI tools and that eighty-two percent of that pasting activity occurs through unmanaged personal accounts, outside any enterprise visibility or oversight [12]. The same report estimates that the average employee performed fourteen pastes per day into non-corporate AI accounts, with at least three containing sensitive data; generative AI had, by that measure, become the single largest channel of corporate-to-personal data movement, larger than personal email, file sharing, or any sanctioned SaaS application [12][13]. Samsung's April 2023 incident, in which three engineers leaked source code, internal meeting transcripts, and chip-fab test sequences into ChatGPT within a twenty-day window of the company allowing the tool, remains the canonical small-scale example of the dynamic; as the CSA AI Safety Initiative's analysis of the shadow AI agent problem documents [23], what has changed since 2023 is that the same behavior is now ambient, automated, and increasingly directed at AI agents with API credentials rather than stateless chat interfaces [14].

Second, the architectural surface of AI inside the enterprise has expanded faster than the responsibility model that governs it. A modern enterprise AI deployment is no longer a single tenant of a hosted LLM. It is a stack of frontier model providers, model gateways, retrieval-augmented generation pipelines anchored to enterprise knowledge bases, vector databases that store embeddings of the documents the enterprise considers most sensitive, and agentic runtimes that execute multi-step tasks across applications using OAuth tokens and machine credentials granted on an ad-hoc basis. Each layer introduces non-human principals — service accounts, API keys, OAuth tokens, MCP server credentials — that authenticate, act, and persist independently of the human identities the IAM team has historically catalogued [15][16]. Industry estimates have put the ratio of non-human identities to human identities in modern enterprises at between twenty-five and fifty to one, with the ratio accelerating as agent deployment scales [15]. Based on the scale and pace of AI agent deployment relative to traditional service account provisioning, AI appears to have become the dominant source of new non-human identity creation in large enterprises, with most of those identities provisioned by application teams or by employees connecting their own AI tools rather than by the identity governance team [15][17][18]. The CSA Shadow Access body of work has spent two years making the case that the accountable owner of access governance must be elevated to cover this expanding surface [17][18].

Third, the regulatory and standards environment has hardened around accountability obligations that the current organizational structure is not designed to meet. The EU AI Act reaches full application for high-risk AI systems on August 2, 2026, and its Article 9 requirement for a documented, lifecycle-spanning risk management system is not satisfiable through the kind of distributed, informal ownership the typical enterprise currently runs [19]. NIST's AI Risk Management Framework Generative AI Profile, finalized as NIST-AI-600-1 in July 2024, treats the "Govern" function as foundational and requires organizations to establish

accountability for AI risk at a level of specificity that few enterprises have committed to in writing [20]. The Digital Operational Resilience Act, in force since January 2025, and the NIS2 Directive, with its October 2026 compliance deadline for in-scope organizations, both impose third-party risk management and incident reporting obligations that depend on the enterprise being able to identify which department owns which AI dependency. The implicit assumption embedded in all of these regimes is that the enterprise has a clear, written, and demonstrable internal accountability map for AI. The 2025 survey evidence is that most enterprises do not [4][5][10].

What follows is an analysis of the three fault lines that have produced the gap, the breach evidence that quantifies the cost of leaving it open, and a control program for closing it.

Anatomy of the Governance Gap

Shadow AI: The Unmanaged Channel

Shadow AI is the contemporary expression of an older phenomenon. Shadow IT, the unsanctioned use of cloud and SaaS tools outside the IT department's catalogue, has been a recognized security category since at least the mid-2010s. Shadow AI shares the same underlying dynamic — employees adopting tools faster than the enterprise can govern them — but its blast radius is different in three important ways.

The first difference is the form of data exposure. Shadow IT historically created exposure through misconfigured services, weak credentials, or third-party breach. Shadow AI creates exposure through voluntary, deliberate disclosure: the employee actively pastes sensitive content into the AI service in order to extract a useful response. The LayerX telemetry data captures this with unusual precision. Whether or not employees are fully aware of the sensitivity of what they paste, the LayerX data shows that twenty-two percent of pasting activity involves PII or PCI data — a pattern that behaves less like an occasional mistake and more like an ambient operational practice the enterprise cannot currently see or control [12]. The same dataset finds that forty percent of file uploads to generative AI tools contain PII or PCI data, and that generative AI now accounts for thirty-two percent of all corporate-to-personal data movement, having surpassed personal email and consumer file-sharing as the leading enterprise exfiltration channel [12][13]. In a control sense, this is closer to a behavioral data-loss problem than a network-security problem, and the controls the enterprise has historically deployed against either category are poorly tuned to it.

The second difference is the absence of a corresponding shadow-AI version of the SaaS Security Posture Management (SSPM) and Cloud Access Security Broker (CASB) toolchain that matured to address shadow IT. CASB and SSPM platforms inventory sanctioned and unsanctioned applications, evaluate their security posture, and apply DLP and access policies at the edge. They were architected, however, against a model in which the unsanctioned tool is a tenant of a discrete cloud service. Shadow AI does not fit that model cleanly.

An employee accessing a personal ChatGPT account through a browser is not interacting with a corporate tenant. The session is unauthenticated against the enterprise, the prompt content lives in a personal account, and the enterprise's CASB sees a browser session to a major SaaS domain that is identical, at the network layer, to a sanctioned access pattern. CSA's own Shadow Access body of work has been explicit that the existing IAM and access-governance toolchain is not adequate to this surface, and that organizations need to treat unauthorized AI access as a discrete category of shadow access requiring tailored controls [17][18][21].

The third difference is the speed and scale of adoption. Shadow IT typically grew at the speed of corporate credit cards and procurement cycles, even at its peak. Shadow AI grows at the speed of browser tabs. The IBM 2025 survey, drawing on roughly 600 breached organizations, found that twenty percent had suffered a breach due to shadow AI in the prior twelve months — a category that did not register at survey-relevant levels in the equivalent 2024 dataset [1][2]. Industry analyses based on browser telemetry and SaaS visibility data have estimated that more than eighty percent of large enterprises now have AI agents and AI tools in active employee use that the security team has not formally onboarded, and that the average enterprise has tens of generative AI services with at least monthly active use among employees but without IT sanction [22]. The CSA AI Safety Initiative's April 2026 analysis of "the shadow AI agent problem" extends the category from chat interfaces to autonomous agents — software entities with API access, persistent credentials, and the ability to chain actions across systems — and notes that the majority of these agents in the enterprise environments studied were created outside the formal application onboarding process [23].

The financial consequence of leaving shadow AI unmanaged is now empirically defined. IBM's 2025 report measured the average cost of a breach involving shadow AI at $4.63 million globally, $670,000 above the comparable cost for breaches not involving shadow AI [1][2]. Customer PII compromise occurred in sixty-five percent of shadow-AI-related breaches, compared with fifty-three percent across breaches in general; intellectual property exposure occurred in forty percent of shadow-AI-related breaches, compared with thirty-three percent at large; and intellectual property exposed through a shadow AI channel carried the highest per-record cost ($178) of any category in the report [1][2]. These numbers should be read against the underlying observation that the controls that would have constrained the exposure — content inspection of outbound AI traffic, identity-bound access to sanctioned tools, classification-aware DLP for AI prompts — are not standard enterprise capabilities in 2026, and the small minority of organizations that have deployed them are correspondingly difficult to compare against the population at large.

Ownership Fragmentation: The Accountability Void

If shadow AI is the channel through which the data leaves the enterprise, ownership fragmentation is the organizational condition that allows it to leave unobserved. The accountability map for enterprise AI in 2026 is, in most large organizations, distributed across at least five executives — the CIO, the CTO, the CISO, the Chief Data Officer or Chief Data and Analytics Officer, and an increasingly common Chief AI Officer — together with the heads of legal, privacy, compliance, internal audit, HR, and the business units that own

particular AI use cases. The 2025 CSA and Google Cloud survey put numbers on the fragmentation. Funding authority for AI security tools is held by the CISO in forty-nine percent of organizations, the CTO in thirty-six percent, the CIO in thirty-three percent, business unit owners in twenty-five percent, and Chief AI Officers in eleven percent, with most organizations reporting overlapping authority across two or more of these roles [4]. Operational responsibility for AI deployment is held by a dedicated AI/ML team in twenty percent of organizations, by IT in nineteen percent, by cross-functional groups in sixteen percent, by the security team itself in thirteen percent, by senior leadership in nine percent, and by data science or analytics in eight percent [4]. Security responsibility for AI specifically falls to the security team in fifty-three percent of organizations, which leaves a substantial minority of cases where AI deployments run with no clear security owner at all [4].

The Boston Consulting Group's broader executive research, drawing on AI strategy interviews with more than a thousand large-enterprise leaders, has found that eighty-five percent of executives consider AI a top priority but only fourteen percent of organizations have clearly defined the roles, responsibilities, and accountability structures required to govern it [5]. State of the CIO survey data from late 2025 found that twenty-four percent of CIOs were uncertain which department is responsible for meeting enterprise AI goals, with ownership "often split across organizations" [24]. Grant Thornton's 2026 AI Impact Survey reported that seventy-eight percent of business executives lacked strong confidence that they could pass an independent AI governance audit within ninety days, a finding that is consistent with the underlying fragmentation [10]. The pattern reported in CIO.com's analysis of the same data captures the operational consequence precisely: responsibility appears distributed while accountability, when tested, is not — and it often compresses to a single point of failure [24].

The compression is the part that matters. In an ordinary year, fragmented ownership presents as friction: AI initiatives move more slowly than the business wants, governance reviews are delayed because three different teams claim authority, and security controls are inconsistently applied. In a breach year, the same fragmentation manifests as an accountability void: regulators, boards, and customers look for the executive who is responsible, and the answer in most current AI deployments is genuinely unclear. The EU AI Act's Article 9 requires providers and deployers of high-risk AI systems to establish a documented, lifecycle-spanning risk management system with clear accountability for each system; Article 26 imposes analogous obligations on deployers [19]. In practice, satisfying these obligations requires designating specific responsible individuals rather than leaving accountability distributed across functions — the regulation does not prescribe a named-individual model by text, but the depth of documentation and oversight it demands renders distributed accountability functionally untenable. NIST's AI 600-1 Generative AI Profile is structurally similar: its "Govern" function requires that organizations document who is accountable for what, and treats undocumented accountability as a deficiency to be remediated [20]. Both regimes assume an organization that can identify, in writing, who owns a given AI risk. Most enterprises in 2026 cannot.

A second consequence of fragmentation, less often noted but operationally significant, is that AI security tooling itself is purchased in a fragmented way. When the CISO funds AI security tools in only forty-nine percent of organizations, the remaining fifty-one percent are buying AI security capability through channels — the CTO, the CIO, business unit budgets, the Chief AI Officer's discretionary spend — that have their own selection criteria, their own integrations with the broader security stack, and their own definitions of "security." The resulting tool sprawl mirrors the ownership sprawl: model gateways purchased by the application team, prompt-injection scanners purchased by the data science team, DLP for AI purchased by the security team, AI governance platforms purchased by legal or compliance, and red-team or evaluation tools purchased by the AI center of excellence. Each of these tools is, in isolation, doing useful work. In aggregate, they produce a posture in which no single dashboard or owner can answer the question "is our AI safe right now?"

Identity and Access: The Technical Blindspot

The third fault line is technical. Existing identity and access management programs were architected around a world in which the principals authenticating to enterprise systems are predominantly humans, with a small set of well-catalogued service accounts and a clearly bounded set of system-to-system integrations. AI has broken that world in three ways simultaneously.

First, AI agents are themselves identity-bearing principals. An MCP server that connects an AI agent to a SaaS tool authenticates as a non-human identity. An AI agent that reads a customer's email, drafts a response, and posts a follow-up to a CRM does so under credentials — typically OAuth tokens or API keys — that persist long after the original task is complete. The Cloud Security Alliance's research on agentic AI and on shadow access has been explicit that these non-human principals require their own lifecycle management, their own principle-of-least-privilege scoping, and their own audit trails [17][21]. Industry surveys put the ratio of non-human identities to humans in modern enterprises in the twenty-five-to-one to fifty-to-one range, with most of those identities provisioned outside the central identity governance program [15][16]. The IBM 2025 report's finding that ninety-seven percent of AI-related breaches occurred at organizations lacking proper AI access controls is, at the technical level, primarily a statement about non-human identity governance [1][3].

Second, the access surface that AI agents touch is broader than the access surface that the same employee would touch through traditional applications. An AI agent that runs on a user's behalf inherits, by default, the entire access scope of that user. As Microsoft's published architecture documentation confirms, Microsoft 365 Copilot in its default configuration can read any document, email, chat message, or SharePoint file that the underlying user account is authorized to access — a scope that varies by licensing tier and has been subject to ongoing refinement following the EchoLeak disclosure; an AI assistant that has been granted access to an enterprise data lake can typically query across the full lake unless explicit row-level or column-level constraints have been configured. The result is a class of permission problem that CSA's Shadow Access publications have called "permission amplification" — the AI agent is not asking for permissions it

does not need; rather, the human user being assisted has accumulated permissions over time, and the AI agent now exercises all of them simultaneously, often combining data from multiple repositories into outputs that the original access controls were not designed to govern [17][18]. The Dark Reading coverage of Microsoft 365 Copilot's no-code agent surface in 2025 documented multiple cases in which low-code-built agents inherited overprovisioned access from their creators and were able to leak customer or financial data without the creator having intended to expose it [25].

Third, the most consequential attack patterns against AI agents in 2025 and 2026 have not been classical credential compromise or privilege escalation. They have been indirect prompt injection through trusted business records, and zero-click exploitation of the data ingestion path. EchoLeak, disclosed as CVE-2025-32711 in June 2025, is the canonical example. An attacker sends an email — a single email — to a Microsoft 365 Copilot user. The email contains, in addition to its visible content, hidden instructions designed to be parsed by the AI when the recipient subsequently asks Copilot to summarize their inbox or related content. The Aim Labs research team that discovered the flaw described a multi-stage chain in which the attacker's instructions evaded Microsoft's Cross-Prompt Injection Attempt classifier, bypassed link-redaction by using reference-style Markdown, exploited Copilot's automatic image fetching, and exfiltrated data through a Teams proxy endpoint whose domain was on Microsoft's own Content Security Policy allow-list [6][7]. The CVSS score assigned by Microsoft was 9.3; the company stated no customers were known to be affected in the wild, and a fix was deployed in the June 2025 Patch Tuesday cycle [6]. The point of the disclosure was not that any one bypass was sophisticated — most of the individual bypasses had been described in research literature — but that the AI agent, by virtue of operating with the user's full access scope and ingesting attacker-controlled content as if it were trusted input, made the entire chain practical against a production service.

AgentFlayer, presented at Black Hat USA 2025 by Zenity Labs, demonstrated that the EchoLeak pattern was not specific to Microsoft. Zenity's working exploit chains compromised OpenAI ChatGPT, Microsoft Copilot Studio, Salesforce Einstein, Google Gemini, and a developer toolchain combining Cursor with the Jira MCP server [8]. The Salesforce Einstein attack hid malicious instructions inside Salesforce records — case entries, lead notes — that the AI agent ingested as part of normal operation; on processing the records, the agent quietly rerouted customer communications to attacker-controlled email addresses, all without triggering Salesforce's built-in defenses [8]. The ForcedLeak disclosure two months later, in September 2025, isolated a closely related flaw in Salesforce Agentforce's Web-to-Lead functionality: an attacker submitting a malicious Web-to-Lead form could exfiltrate CRM data through the agent's downstream actions, with the issue receiving a CVSS score of 9.4 from researcher Noma Security and a patch from Salesforce that involved enforcing Trusted URLs and reclaiming an expired domain attackers could otherwise have repurposed [9]. None of these flaws were the work of advanced persistent threat actors. They were demonstrated by commercial research labs working within responsible-disclosure timelines, against the default configurations of the dominant enterprise AI agent platforms.

The conclusion that follows is that the AI agent layer is, at the moment, an under-governed identity surface with an under-defended ingestion path — where "adequate" governance would mean least-privilege access scoping, content-trust boundaries, and audit logging at the agent layer, none of which are deployed at scale. The standard enterprise response — patch the vendor flaw, wait for the next one — is necessary but not sufficient. The deeper problem is that the controls that would have constrained the blast radius of these attacks (least-privilege access scoping on the agent, content-trust boundaries between human-authored and machine-ingested input, audit logging of agent actions, output filtering for sensitive content) are not deployed at scale, and are not in most organizations owned by a single accountable function.

Quantifying the Breach Blindspot

The 2025 breach data is the first dataset large and consistent enough to support an empirical view of what the AI Security Governance Gap costs. The headline figures from IBM's 2025 Cost of a Data Breach report, conducted by Ponemon Institute on IBM's behalf and based on incidents at 600 organizations across 16 countries between March 2024 and February 2025, are striking on three dimensions [1][2]. IBM Security, which commissioned the research, offers commercial breach response and AI security services; the dataset is nonetheless widely cited across the security research community as the most comprehensive annual breach-cost study available.

On the dimension of frequency, AI is now a measurable category in the breach data. Thirteen percent of organizations reported a breach involving an AI model or AI application directly, with an additional eight percent unsure whether they had been compromised in that way; twenty percent reported a breach involving shadow AI; and sixteen percent of all breaches involved attackers using AI tools, predominantly for AI-generated phishing (thirty-seven percent of AI-enabled attacks) and deepfake impersonation (thirty-five percent) [1][2][3]. The combined picture is that, by mid-2025, AI was either the asset under attack, the channel of attack, or both, in a substantial minority of all enterprise breaches. None of these categories existed in the breach taxonomy two years earlier.

On the dimension of cost, the absence of governance is correlated with materially larger losses. Breaches involving shadow AI averaged $4.63 million globally, compared with $3.96 million for breaches not involving shadow AI — a $670,000 differential [1][2]. The global average breach cost across all categories was $4.44 million, down for the first time in five years from the prior year's $4.88 million, with the decline attributed in significant part to faster breach detection and containment among organizations using AI in their security operations; the cost differential between shadow AI breaches and other breaches widens against that backdrop [1][2]. Customer PII compromise occurred in sixty-five percent of shadow-AI breaches versus fifty-three percent across breaches in general; intellectual property exposure occurred in forty percent versus thirty-three percent; and IP exposed through shadow AI channels carried the highest per-record cost ($178) of any category in the report [1][2].

On the dimension of root-cause attribution, the IBM data is unusually direct about where the gap is. Ninety-seven percent of organizations reporting an AI-related breach lacked proper AI access controls, sixty-three percent of breached organizations had no AI governance policy or were still developing one, and only thirty-four percent of those with a policy performed regular audits for unauthorized AI use [1][2][3]. The ninety-seven percent figure is worth dwelling on. It is rare for a single attribute to appear in nearly every breach within a category. Access control failures specific to AI principals — overprovisioned agent identities, missing scoping on OAuth tokens granted to AI tools, absence of revocation mechanisms when a tool is no longer in use — are not, in this dataset, one risk factor among many. They are nearly universal in the breached population.

A useful counterpoint sits in the same report. Organizations using AI defensively, in their security operations, were measured to cut their breach lifecycle by eighty days on average and save roughly $1.9 million in associated costs [1][2]. The same technology that creates the governance gap on the consumption side closes it on the defense side, but only for organizations that have actually deployed it in a structured way. The 2025 CSA and Google Cloud survey found that ninety percent of organizations were testing or planning AI for security use cases [4], which is encouraging at the adoption level; it does not change the fact that the protective benefit accrues disproportionately to the minority of organizations that already have governance in place.

Three further pieces of incident-level evidence, drawn from public disclosures across 2025 and the first months of 2026, illustrate how the breach pattern actually unfolds in practice. The first is the EchoLeak disclosure and patch cycle, described in the previous section: a single email exfiltrating arbitrary Copilot-accessible data, with no user interaction, through a chain of allow-listed delivery paths that Microsoft itself had configured [6][7]. The second is the AgentFlayer family of zero-click exploit chains demonstrated by Zenity Labs against the leading enterprise AI agent platforms, with the Salesforce Einstein attack accomplishing data exfiltration and the silent rerouting of customer communications through indirect prompt injection embedded in business records [8]. The third is the ForcedLeak Salesforce Agentforce flaw, where the trust boundary between an external Web-to-Lead submission and an internal AI agent was, in default configurations, weaker than the trust boundary between an external email and an internal user [9]. These three disclosures do not exhaust the public 2025–2026 AI agent incident catalogue, but together they establish that the technical preconditions for the IBM survey's findings are present in production today, at the scale of the dominant enterprise AI platforms.

The Regulatory and Standards Inflection

The regulatory environment around enterprise AI has been moving faster than the operational governance environment. Three regulatory developments will, between now and the end of 2026, push the accountability question from a recommended best practice into a binding obligation for most large organizations.

The EU AI Act reached full application in stages beginning in August 2024, with the rules for high-risk AI systems entering binding force on August 2, 2026. Article 9 requires that providers and, by extension, deployers of high-risk AI systems establish, implement, document, and maintain a lifecycle-spanning risk management system that identifies known and foreseeable risks, estimates risk exposure during intended use and foreseeable misuse, evaluates emerging risks from post-market monitoring, and implements targeted mitigations [19]. The combined effect of Articles 9 through 17 — covering risk management, data governance, technical documentation, automatic event logging, transparency, and human oversight — is that any organization deploying an AI system in a high-risk domain (which under Annex III includes employment, education, essential services, law enforcement, and certain financial-services and critical-infrastructure use cases) will need to demonstrate, in writing, who owns each control and who is accountable for each system. Article 26 imposes analogous obligations on deployers. The structural implication is that accountability must be formalized rather than assumed.

NIST's AI Risk Management Framework, with the Generative AI Profile published as NIST-AI-600-1 in July 2024, is voluntary in formal terms but has rapidly become the default reference standard in US federal procurement and in financial-sector supervisory practice [20]. The framework's four functions — Govern, Map, Measure, Manage — are structured so that Govern is foundational. Roles, responsibilities, accountability structures, escalation paths, policy ownership, and documentation are the precondition for the rest of the framework to operate. Organizations that have not done the Govern work cannot meaningfully Measure or Manage; the framework is explicit on that ordering [20]. The 2025 surveys make clear that most enterprises are starting from a position in which the Govern function is the weakest.

In parallel, financial and critical-services sectors have moved on operational-resilience obligations that have spillover effects for AI governance. The EU's Digital Operational Resilience Act has been in force for regulated financial entities since January 17, 2025, and on November 18, 2025 the European Supervisory Authorities published the first list of nineteen Critical ICT Third-Party Providers, which includes the major hyperscalers that host most enterprise AI workloads. NIS2 reaches its compliance deadline for in-scope organizations in October 2026 and imposes third-party risk management and incident reporting obligations on a substantially broader set of essential and important entities. Both regimes assume that the regulated entity can identify, on a per-service basis, which third-party AI dependencies it has, who owns them internally, and what the failure modes are. The internal mapping work this requires is not trivial in an environment where the AI service catalogue is itself incomplete.

For United States enterprises, the patchwork is messier but moving in the same direction. The SEC's cybersecurity disclosure rules, in force since 2024, require timely reporting of material cyber incidents and have begun to be applied to AI-related events; the New York Department of Financial Services' guidance on AI-related cybersecurity risks, published in October 2024, sets specific expectations around third-party risk management, multi-factor authentication, and the cybersecurity governance of AI deployments at regulated entities. State attorneys general have begun to use existing consumer protection and data privacy

authorities to pursue enforcement actions against AI-related data exposures. The cumulative regulatory effect, on both sides of the Atlantic, is that "we don't know who owns this" is moving from an internal operational embarrassment to a regulatory deficiency.

Why Existing Controls Don't Close the Gap

Many of the controls that the AI governance gap requires already exist in the enterprise security toolkit, applied to other categories of risk. The reason they have not closed the gap is that they have not been adapted to the specifics of how AI consumes data, who acts on whose behalf, and what an AI principal's identity actually is. Three categories of control illustrate the adaptation challenge.

Data loss prevention is the closest existing capability to the shadow-AI exfiltration channel, but the typical DLP deployment was tuned to a different threat model. Classical DLP looks for sensitive content (credit card numbers, social security numbers, source code patterns, regulated health information) traversing network egress paths or storage interfaces, with the assumption that content matching a policy and leaving an authorized boundary represents potential exfiltration. AI traffic does not fit cleanly. The content an employee pastes into an AI prompt is often not regex-detectable in the same way that classical PII is. The destination is, at the network layer, indistinguishable from sanctioned SaaS access. And the content is being moved through a browser session under the user's own credentials, which is exactly the pattern DLP policies were configured to allow. A handful of vendors now offer AI-specific DLP that classifies and filters prompt content at the browser or proxy layer, but coverage at the enterprise level remains limited, and security practitioners have documented that false-positive rates on creative and analytical work frequently lead organizations to loosen AI DLP policies under user pressure, limiting effective coverage in practice.

SaaS Security Posture Management is the closest existing capability to the unsanctioned-AI-tool inventory problem, but SSPM was architected around tenanted relationships between the enterprise and a SaaS provider. SSPM excels at evaluating the configuration of a sanctioned tenant — checking for over-permissive sharing, weak authentication policies, or misconfigured integrations — and at discovering unsanctioned tenants through OAuth grant analysis. It does not, in its standard form, address the case where an employee is using a personal AI account from a corporate device with no enterprise tenant of any kind. CASB platforms can layer additional visibility through proxy-based session inspection, but the effectiveness depends on the depth of decryption and the willingness of the organization to deploy proxy-level controls broadly enough to cover the actual exfiltration pattern. The CSA Shadow Access body of work has been explicit that the shadow access category — including shadow AI access — requires controls that integrate browser, identity, network, and content perspectives in ways that no single existing toolclass covers [17][18].

Identity governance is the closest existing capability to the non-human-identity problem that AI introduces. Modern Identity Governance and Administration platforms can inventory service accounts, manage entitlement reviews, enforce least-privilege scoping, and rotate credentials. The challenge is that the AI

principal landscape — OAuth tokens granted to AI tools by individual employees, API keys created in developer consoles for AI integration projects, MCP server credentials provisioned by application teams, agent runtimes that authenticate as ephemeral principals — does not map cleanly into the entitlement-and-review model that IGA platforms were designed around. The CSA Confronting Shadow Access publication and the related Zero Trust and IAM guidance set out the case for extending IGA to cover these principals, but operational deployment has lagged the framework guidance [17][21]. The IBM 2025 finding that ninety-seven percent of AI-related breaches involved organizations lacking proper AI access controls is, in practical terms, a finding that the IGA gap on non-human identities has not yet been closed [1][3].

The cross-cutting reason these controls have not closed the gap, even where they have been individually deployed, is that they were purchased and operated by different parts of the organization. DLP is typically owned by the security team. SSPM and CASB are often owned by the cloud security team. Identity governance is owned by an identity team that may report to the CIO rather than the CISO. AI tooling is increasingly owned by an AI center of excellence that reports to the Chief AI Officer or the Chief Data Officer. Each of these owners has its own roadmap, its own KPIs, and its own definition of success. The owner of the AI governance program — to the extent there is one — is rarely empowered to direct the others. Ownership fragmentation, having produced the gap, also blocks its closure.

A Control Program Aligned to CSA Frameworks

CSA's existing research catalogue provides most of the components a comprehensive AI governance program requires. What follows is a control program organized around three CSA frameworks — the AI Controls Matrix (AICM), the MAESTRO agentic AI threat modeling framework, and the Shadow Access and Zero Trust IAM body of work — together with the cross-cutting accountability guidance in CSA's AI Organizational Responsibilities publications. CSA's frameworks complement, and are designed to integrate with, related industry standards including ISO/IEC 42001 for AI management systems and MITRE ATLAS for adversarial threat modeling; organizations already invested in those frameworks can use AICM and MAESTRO as extensions rather than replacements.

The AI Controls Matrix, released in July 2025, provides the foundational control taxonomy. AICM consists of 243 control objectives distributed across eighteen security domains, ranging from traditional categories like Identity and Access Management and Data Security and Privacy Lifecycle Management to AI-specific domains like Model Security and Supply Chain Management, Transparency, and Accountability [26][27]. The framework's Shared Security Responsibility Model assigns control ownership across five actors — Cloud Service Providers, Model Providers, Orchestrated Service Providers, Application Providers, and AI Customers — which gives the enterprise a reference for which controls it can rely on its suppliers to provide

and which it must implement directly [26][27]. The August 2025 CSA blog on strategic implementation of AICM provides a CISO-oriented walkthrough of how to phase deployment of the controls, starting with the highest-leverage subset and expanding [28].

Within AICM, the controls most directly responsive to the governance gap fall in three domains. Identity and Access Management controls address the non-human identity surface, including provisioning, scoping, lifecycle management, and revocation of agent and AI tool credentials. Governance, Risk Management, and Compliance controls address the accountability architecture, including documented role assignments, policy ownership, and escalation paths. Data Security and Privacy Lifecycle Management controls address the shadow AI exfiltration channel, including classification, retention, and exfiltration prevention for content entering and leaving AI systems. The eighteen-domain breadth of AICM is itself a corrective to the fragmentation pattern: the matrix forces the question "who owns this control?" to be asked and answered for every cell, which surfaces ownership voids before they become breach attribution problems.

MAESTRO — the Multi-Agent Environment, Security, Threat, Risk, and Outcome framework released by CSA in February 2025 — extends the control discussion into agentic AI threat modeling [29]. MAESTRO is structured around a seven-layer reference architecture that decomposes an agentic AI system into discrete analysis surfaces: foundation models, data operations, agent frameworks, deployment infrastructure, a vertical security and compliance layer, the agent ecosystem, and an evaluation and observability layer. The framework adapts established threat-modeling approaches (STRIDE, PASTA, LINDDUN) for the specifics of autonomous agents, including the prompt-injection, indirect-prompt-injection, and tool-misuse attack patterns that EchoLeak, AgentFlayer, and ForcedLeak exemplify [29]. Applied to the breach evidence, MAESTRO is useful in two specific ways. First, the layered decomposition forces explicit modeling of the trust boundary between human-authored input and machine-ingested content, which is the boundary at which most of the zero-click attacks crossed in 2025. Second, the evaluation and observability layer prompts the question of which monitoring and logging exists at the agent layer, which in most current deployments is a substantial gap.

The CSA Shadow Access and Zero Trust IAM body of work — including the Confronting Shadow Access Risks publication, the Shadow Access and AI publication, and the related blog and working-group output — provides the identity-and-access architecture for closing the non-human identity surface that the breach data identifies as the dominant root cause [17][18][21][30]. The Zero Trust principle of "never trust, always verify" applies as directly to AI principals as it does to human ones, and arguably more so given that AI principals are more numerous, more dynamic, and more capable of exercising broad access scopes than humans are. The practical implication of the Shadow Access guidance for AI is that every AI tool integration, every MCP server, every agent runtime, and every model gateway should be inventoried, owned, scoped to least-privilege access against the specific data and capability it requires, monitored at the audit-log level, and subject to lifecycle deprovisioning when no longer in use. None of these requirements is technically novel; what is novel is the scale at which they must be applied, and the fact that the responsible owner in most current organizations is undefined.

Cutting across all of these frameworks, CSA's AI Organizational Responsibilities publications — covering governance, risk management, compliance, and cultural aspects; core security responsibilities; and AI tools and applications — provide the accountability layer that the technical controls require to be effective [31]. The publications set out role definitions for the executive functions that need to be involved in AI governance (CISO, CIO, CTO, Chief AI Officer, Chief Data Officer, Chief Risk Officer, Chief Privacy Officer, Chief Legal Officer, business unit owners), specify decision rights and escalation paths, and provide templates for the documented accountability assignments that the EU AI Act, NIST AI RMF, and emerging US regulatory frameworks now require. The structural recommendation in this body of work — that organizations consolidate AI accountability under a single named executive sponsor, with cross-functional working-group support, rather than allowing it to remain distributed by default — is the most direct organizational answer to the ownership fragmentation that the survey data describes.

A practical sequence for an enterprise adopting this program is to begin with the AI Organizational Responsibilities accountability work (because the governance program will fail in execution without a single owner), proceed to the AICM control taxonomy (because it surfaces the gaps systematically), apply MAESTRO threat modeling to the highest-risk agentic AI deployments first (because the breach evidence concentrates there), and extend the Shadow Access and Zero Trust IAM controls to the non-human identity surface as a continuous program (because the surface itself is growing). The same sequence supports the EU AI Act's Article 9 obligations on lifecycle risk management, NIST AI RMF's Govern function, and the operational resilience requirements emerging under DORA and NIS2.

Recommendations for Security Leaders

Three categories of action are available to security leaders in the current planning cycle. They differ in horizon and in the kind of organizational support they require.

The immediate-action category, for the next sixty to ninety days, focuses on visibility and accountability. Conducting a shadow AI discovery exercise — combining browser telemetry, OAuth grant audits, and employee survey — establishes the baseline against which controls will be measured. Identifying the executive accountable for the enterprise AI security program, and documenting that accountability in writing with the explicit endorsement of the CEO or board risk committee, addresses the ownership fragmentation at the level it has to be addressed. Inventorying the existing AI tool catalogue, including agents, model gateways, MCP servers, and AI features inside SaaS applications, surfaces the non-human identity surface that needs to be governed. For most mid-sized enterprises, each of these actions is achievable within a quarter without procuring new tooling; larger or more complex organizations should expect proportionally more time for the shadow AI discovery and executive alignment steps.

The short-term mitigation category, for the next six to twelve months, focuses on closing the most acute identified gaps. Extending the existing identity governance program to cover AI-related non-human principals, with provisioning, entitlement review, and lifecycle deprovisioning aligned to the patterns CSA's Shadow Access guidance describes, addresses the technical root cause that the IBM 2025 data identifies as nearly universal in AI breaches. Deploying AI-aware data-loss prevention at the browser or proxy layer, with classification policies tuned to the prompt-content patterns that LayerX and similar telemetry research has documented, narrows the shadow AI exfiltration channel. Applying MAESTRO threat modeling to the highest-business-risk agentic AI deployments — typically those with broad data access scope, multi-step task execution authority, or external content ingestion — generates the prioritized control list for the deployments where exposure is largest. Establishing logging and observability at the agent action layer addresses the evaluation gap that allows current zero-click attacks to operate without detection. Implementing the AICM control objectives in the Identity and Access Management, Governance Risk Compliance, and Data Security and Privacy Lifecycle Management domains, with explicit ownership assigned to named functions, establishes the foundation against which further controls can layer.

The strategic-consideration category, on a twelve-to-thirty-six-month horizon, focuses on the structural changes that prevent the gap from re-emerging. Consolidating AI security accountability under a single named executive sponsor, with budgetary authority for AI security tooling, addresses the funding fragmentation that has produced overlapping tool deployments. Integrating AI governance into board-level risk reporting, with documented metrics and named accountability, brings the AI risk discussion into the same governance regime as other enterprise risk categories. Aligning organizational AI policy with the EU AI Act, NIST AI RMF, DORA, and NIS2 obligations creates a single internal compliance fabric that can absorb future regulatory developments without re-architecting the program each time. Investing in AI security skill building across the security organization closes the skills-gap component that the CSA and Google Cloud 2025 survey identified as the top hurdle for sixty-one percent of respondents [4]. Finally, embedding AI governance into the procurement and vendor risk management process — making AICM Shared Security Responsibility Model attestations and equivalent vendor commitments part of standard enterprise contracting — closes the supply chain dimension of the gap that the AICM framework explicitly addresses [26][27].

These recommendations have appeared in governance guidance for several years. What is new is the breach evidence that converts them from good practice into measurable loss mitigation, and the regulatory environment that converts them from voluntary to auditable. The organizations that close the gap in this planning cycle will be operating against a different risk profile than those that wait for the next disclosure to force their hand.

Conclusion

The AI Security Governance Gap is no longer a forward-looking concern. The IBM 2025 Cost of a Data Breach data has converted it into a measurable line item in enterprise loss exposure: a $670,000 differential per breach for organizations with shadow AI activity, a nearly universal access-control failure pattern in AI-related breaches, and a sixty-three percent rate of breached organizations without an AI governance policy [1][2]. The CSA and Google Cloud State of AI Security and Governance survey has shown that the structural cause is governance fragmentation rather than tool absence: most large enterprises now have at least some AI security tooling, but few have established the unified accountability under which that tooling could operate as a coherent program [4]. The breach disclosures of 2025 and early 2026 — EchoLeak, AgentFlayer, ForcedLeak, and the broader pattern of zero-click and indirect prompt injection against the dominant enterprise AI platforms — have shown that the technical surface is real and that production exploit chains are achievable against default configurations by mainstream commercial research labs [6][7][8][9].

The regulatory environment is closing on the same problem from a different direction. The EU AI Act's August 2, 2026 high-risk system obligations, NIST AI RMF's Govern function, DORA's third-party operational resilience requirements, and NIS2's incident reporting and supply chain obligations all assume an enterprise that has, in writing, identified who owns each AI risk and who is accountable when a control fails. The current survey evidence is that most enterprises have not yet built that map. The August 2, 2026 EU deadline is approximately ten weeks from the date of this paper.

CSA's existing research catalogue — the AI Controls Matrix, MAESTRO, Confronting Shadow Access Risks, Shadow Access and AI, Zero Trust IAM guidance, and the AI Organizational Responsibilities publications — provides most of the components a comprehensive program requires. The work that remains is organizational and is, in nearly every case, less expensive than the breach it would prevent. The starting point is the question that the survey data, the breach data, and the regulatory environment all converge on: who, by name, in your organization is accountable for AI security, and can that person produce, this quarter, a documented map of the AI principals, dependencies, and controls under their stewardship? Where the answer is yes — where a named executive can produce that map within a quarter — the program is recoverable. Where it takes longer than a quarter to answer the question at all, the urgency is what the breach data says it is.

CSA Resource Alignment

The work in this paper is anchored in, and directly supports, the following Cloud Security Alliance frameworks and publications. Practitioners closing the AI Security Governance Gap should treat these as the operating reference set.

The **AI Controls Matrix (AICM)** is the primary control taxonomy for AI-specific risk, with 243 control objectives across eighteen domains and a Shared Security Responsibility Model that assigns ownership across cloud, model, orchestration, application, and customer roles [26][27][28]. Implementation should begin with the Identity and Access Management, Governance Risk and Compliance, and Data Security and Privacy Lifecycle Management domains as the most direct responses to the shadow AI, ownership fragmentation, and non-human identity surfaces analyzed here.

The **MAESTRO Agentic AI Threat Modeling Framework** decomposes agentic AI systems into seven layers and provides the methodology for systematic threat analysis at each layer [29]. MAESTRO should be applied to the highest-risk agentic deployments before broader rollout — typically those with broad data access, multi-step task execution, or external content ingestion — and should be integrated into the development lifecycle for new AI agent capabilities.

The **Confronting Shadow Access Risks** and **Shadow Access and AI** publications, together with the broader Zero Trust IAM body of CSA work, provide the identity and access architecture for governing the non-human principals (agents, MCP servers, model gateways, OAuth-token-bearing AI tools) that AI has introduced into the enterprise [17][18][21][30]. These should drive extension of existing identity governance programs to cover the AI-specific principal landscape.

The **AI Organizational Responsibilities** publication suite — covering governance, risk management, compliance, and cultural aspects; core security responsibilities; and AI tools and applications — provides the executive-level accountability framework that translates the technical controls into organizational practice [31]. These should drive the consolidation of AI security accountability under a single named executive sponsor and the documented assignment of decision rights across the C-suite.

The **AICM Implementation and Auditing Guidelines** publications, segmented by actor role (Cloud Service Provider, Model Provider, Orchestrated Service Provider, Application Provider, AI Customer), translate the AICM controls into role-specific implementation and assessment guidance. Enterprises should use the AI Customer guidance as their primary implementation reference and use the other role-specific guidance to inform vendor security expectations and procurement requirements.

The CSA AI Safety Initiative's ongoing work on agent identity, MCP security, and AI supply chain — visible in the published research on the Agentic Trust Framework and the Shadow AI Agent Problem in Enterprise Environments [23] — provides the current view of the evolving threat surface and should be tracked by security organizations as the operating frontier of AI governance practice.

References

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

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

[3] IBM Newsroom. "IBM Report: 13% Of Organizations Reported Breaches Of AI Models Or Applications, 97% Of Which Reported Lacking Proper AI Access Controls." IBM, July 30, 2025.

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

[5] Boston Consulting Group. "Closing the AI Impact Gap." BCG, 2025.

[6] Dark Reading. "Researchers Detail Zero-Click Copilot Exploit 'EchoLeak'." Dark Reading, June 2025.

[7] The Hacker News. "Zero-Click AI Vulnerability Exposes Microsoft 365 Copilot Data Without User Interaction." The Hacker News, June 2025.

[8] PR Newswire / Zenity Labs. "Zenity Labs Exposes Widespread 'AgentFlayer' Vulnerabilities Allowing Silent Hijacking of Major Enterprise AI Agents Circumventing Human Oversight." PR Newswire, August 2025.

[9] The Hacker News. "Salesforce Patches Critical ForcedLeak Bug Exposing CRM Data via AI Prompt Injection." The Hacker News, September 2025.

[10] Grant Thornton. "2026 AI Impact Survey Report." Grant Thornton, 2026.

[11] McKinsey. "The State of AI: Global Survey 2025." McKinsey & Company, 2025.

[12] LayerX. "The LayerX Enterprise AI & SaaS Data Security Report 2025." LayerX, 2025.

[13] The Hacker News. "New Research: AI Is Already the #1 Data Exfiltration Channel in the Enterprise." The Hacker News, October 2025.

[14] TechCrunch. "Samsung bans use of generative AI tools like ChatGPT after April internal data leak." TechCrunch, May 2, 2023.

[15] Token Security. "The Ultimate Non-Human Identity Security Guide." Token Security, 2026.

[16] The Hacker News. "The Non-Human Identity Crisis: Why Your Machine Identities Are Your Biggest Governance Gap." The Hacker News, May 2026.

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

[18] Cloud Security Alliance. "Zero Trust & Identity and Access Management (IAM): Mitigating Shadow Access." Cloud Security Alliance Blog, May 10, 2024.

[19] EU Artificial Intelligence Act. "Article 9: Risk Management System." EU AI Act (Regulation 2024/1689).

[20] NIST. "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)." National Institute of Standards and Technology, July 2024.

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

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

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

[24] CIO.com. "AI is spreading decision-making, but not accountability." CIO, 2026.

[25] Dark Reading. "Copilot's No-Code AI Agents Liable to Leak Company Data." Dark Reading, 2025.

[26] Cloud Security Alliance. "AI Controls Matrix." Cloud Security Alliance, July 2025.

[27] Cloud Security Alliance. "Introducing the CSA AI Controls Matrix: A Comprehensive Framework for Trustworthy AI." Cloud Security Alliance Blog, July 10, 2025.

[28] Cloud Security Alliance. "Strategic Implementation of the CSA AI Controls Matrix: A CISO's Guide to Trustworthy AI Governance." Cloud Security Alliance Blog, August 8, 2025.

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

[30] Cloud Security Alliance. "The Agentic Trust Framework: Zero Trust Governance for AI Agents." Cloud Security Alliance Blog, February 2, 2026.

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

