International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 514-524
https://doi.org/10.70917/ijcisim-2026-5637
Article

Shadow AI in the Enterprise: A Governance Framework for Transitioning from Uncontrolled Experimentation to Secure, Accountable AI Adoption

Abhipray Mirke[1], Sushmita Dey Banik[2]

[1]Senior Manager - Finance & CISO, India. Email ID - abhipray.mirke@shell.com
[2]GRC PRODUCT SPECIALIST, India. Email ID - sushmitabanik83@gmail.com

Abstract: The rapid, low-friction availability of generative artificial intelligence (AI) has produced a class of unmanaged enterprise technology risk referred to as Shadow AI which means the use of AI tools, models, application programming interfaces (APIs), browser extensions, copilots, and increasingly autonomous agents for business purposes without formal approval, security review, or governance oversight. This paper synthesizes industry survey data, regulatory texts, and documented real-world incidents to characterize the scale, drivers, and risk surface of Shadow AI, and to propose a structured governance framework for converting unsanctioned use into accountable Governed AI. Using a qualitative document-analysis methodology applied to seventeen primary sources spanning vendor research, security standards bodies, and regulatory instruments, the paper finds that Shadow AI is driven primarily by a productivity gap rather than malicious intent, that organizations with high Shadow AI exposure incur materially higher data-breach costs, and that the emergence of autonomous AI agents constitutes a distinct and escalating governance challenge that traditional acceptable-use policies cannot adequately address. The paper presents a ten-principle governance model, a risk-tiering structure, an agent-identity control baseline, and a four-phase implementation roadmap, concluding that prohibition-based strategies are largely ineffective and that visibility, safe enablement, and embedded controls produce more durable risk reduction.

Keywords: Shadow AI; AI governance; enterprise risk management; generative AI; autonomous agents; data privacy compliance; cybersecurity policy

1. Introduction

Artificial intelligence has moved from a specialist enterprise capability to a freely accessible consumer technology in a remarkably short period. Generative AI tools can now be reached through a web browser, a free account, or a single plug-in installation, removing the procurement, infrastructure, and engineering barriers that historically constrained how quickly new technology entered the workplace [1]. This frictionless accessibility has produced a phenomenon the industry terms "Shadow AI": the use of AI models, copilots, automation scripts, APIs, and AI-enabled software-as-a-service (SaaS) features for business purposes without formal approval, visibility, security review, or lifecycle governance from the employing organisation [2].

Shadow AI is conceptually related to "Shadow IT"- the long-standing problem of employees adopting unsanctioned hardware or software - but its risk surface is substantially larger. Whereas Shadow IT tools primarily store or transmit data, AI systems can infer, generate, automate, recommend, write executable code, connect to enterprise systems through APIs, and, in their most advanced form, act autonomously on behalf of a user [2]. This expansion from passive data handling to active decision-making changes the risk calculus for governance, security, privacy, and legal functions simultaneously.

Industry survey evidence indicates that Shadow AI is not a marginal phenomenon. The Microsoft and LinkedIn 2024 Work Trend Index found that 75% of knowledge workers were already using AI at work and that 78% of AI users were bringing personal AI tools into the workplace, creating a measurable gap between employee adoption and organisational control [1]. Gartner has projected that a significant share of enterprises will experience a security or compliance incident linked to unauthorised AI use before the end of the decade, while 69% of surveyed cybersecurity leaders already suspect or have evidence of prohibited generative AI use inside their organisations [3], [4]. KPMG's 2025 global study further revealed that only 41% of employees reported that their organisation had a documented policy governing generative AI use [5].

The purpose of this paper is threefold. First, it characterises the definitional boundaries, growth drivers, and cross-functional manifestations of Shadow AI. Second, it analyses the associated risk landscape, drawing on the OWASP Top 10 for Large Language Model (LLM) Applications and on documented breach-cost data, to establish why Shadow AI represents a measurable and growing source of enterprise loss [9], [11]. Third, it proposes a structured governance response comprising guiding principles, an operating model, a technology architecture, and a four-phase implementation roadmap, with particular attention to the governance of autonomous AI agents as a distinct and more dangerous evolution of the problem [2], [6].

The central argument is that prohibition alone is an insufficient and frequently counter-productive response. Banning AI tools tends to drive usage underground rather than eliminating it, because the underlying productivity pressure that motivates adoption does not disappear when access is blocked [5]. The paper therefore argues for a strategy of visibility before restriction and safe enablement, in which governance is embedded into employee workflows rather than imposed as a separate, slow-moving approval process [2].

2. Methodology

This paper adopts a qualitative, document-based research design appropriate for a fast-moving, practice-oriented governance topic for which controlled experimentation is neither feasible nor ethical. The methodology comprises four stages: source identification, source triangulation, thematic synthesis, and framework construction.

A. Source Identification

Primary sources were drawn from four categories: (i) large-sample industry workplace surveys, including the Microsoft and LinkedIn Work Trend Index, the KPMG global AI trust study, and the McKinsey State of AI survey; (ii) security and risk research published by recognised standards bodies, including OWASP, NIST, ISO/IEC, and IBM Security; (iii) regulatory and legal primary texts, including the EU Artificial Intelligence Act, the General Data Protection Regulation (GDPR), and India's Digital Personal Data Protection Act, 2023; and (iv) documented real-world incidents, including the Samsung source-code exposure episode and the Air Canada chatbot liability tribunal decision. Sources were restricted to those published between 2023 and 2025 to ensure relevance to the current generative AI and agentic AI landscape.

B. Source Triangulation

Where a quantitative claim appears in more than one secondary report of the same underlying study - for example, IBM Cost of a Data Breach findings reported independently by Cybersecurity Dive and VentureBeat - both the primary report and at least one independent secondary report were retained and cross-checked for consistency before the figure was incorporated into the analysis. This reduces the risk that a single outlet's interpretation introduces distortion into the synthesis.

C. Thematic Synthesis

Sources were coded against six analytic categories derived inductively from an initial reading of the literature: (1) adoption drivers, (2) functional manifestation across business units, (3) technical and security risk, (4) regulatory and legal exposure, (5) economic and breach-cost impact, and (6) governance and control mechanisms. Recurring

themes within each category were consolidated into the Results and Discussion section, and illustrative statistics were extracted for quantitative summary in Table I and Fig. 1.

D. Framework Construction

The governance framework was constructed by mapping the risk themes identified during thematic synthesis against established control domains from the NIST AI Risk Management Framework, the OWASP LLM Top 10, and ISO/IEC 42001, and by incorporating identity-governance principles adapted from established privileged-access-management practice to the novel case of autonomous AI agents. This mapping exercise produced a ten-principle governance model, a five-level agent privilege model, and a four-phase implementation roadmap.

E. Limitations

As a document-analysis study, this paper does not generate new primary survey or breach data; it relies on the methodological rigour of the underlying cited studies, several of which use self-reported, vendor-sponsored survey instruments that may carry social-desirability or selection bias. Statistics should therefore be read as directionally indicative of an industry-wide pattern rather than as precise population estimates.

3. Results and Discussion

This section presents the principal findings of the synthesis, organised into five subsections: the adoption-governance gap, the functional risk landscape, the economics of Shadow AI, the emergent risk of autonomous agents, and the regulatory compliance gap.

A. The Adoption-Governance Gap

Across the triangulated survey sources, a consistent pattern emerges: employee adoption of AI consistently and substantially outpaces formal enterprise governance. Seventy-five percent of knowledge workers reported using AI at work, and 78% of those AI users were bringing their own tools into the workplace independently of any enterprise programme, creating a measurable gap between adoption and organisational control. Sixty-nine percent of cybersecurity leaders surveyed either suspected or had confirmed evidence of prohibited generative-AI use within their organisations, while only 41% of employees reported that their organisation had a documented policy governing generative AI use. Separately, 62% of surveyed organisations reported experimenting with or having deployed AI agents, even though most respondents described their organisations as early in scaling AI enterprise-wide. Fig. 1a visualises this adoption-governance gap.

This pattern is structurally different from earlier Shadow IT waves because no infrastructure, procurement cycle, or specialised technical skill is required for an employee to begin using a capable AI system. A browser, a personal account, or a single plug-in installation is sufficient, meaning the adoption curve moves faster than any procurement or policy review process can track.

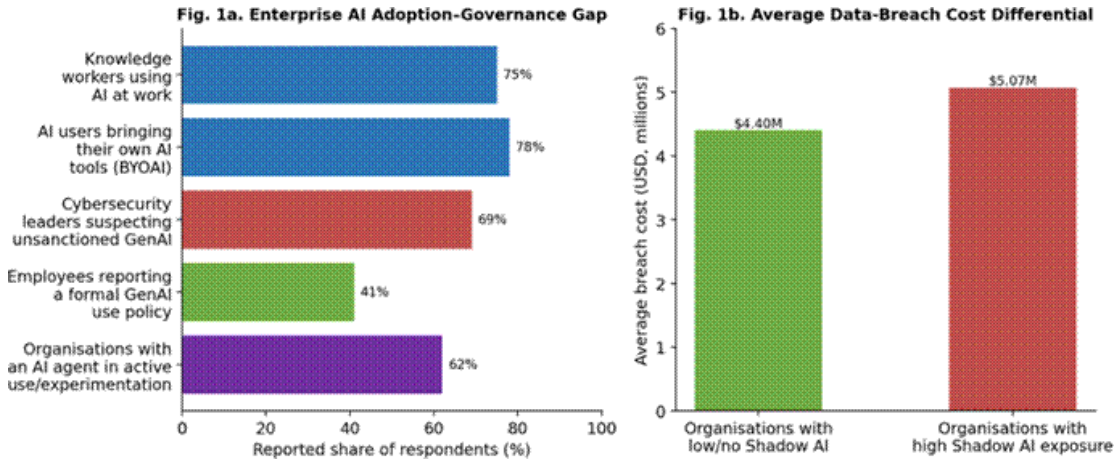

Fig. 1. (a) Enterprise AI adoption–governance gap across five reported survey indicators; (b) average data-breach cost differential between low and high Shadow AI exposure organisations.

B. Functional Risk Landscape

Thematic coding of the source material identified eight cross-functional risk themes, summarised in Table I. Each risk theme recurs across multiple independent sources, increasing confidence that the pattern reflects a genuine, broad-based exposure rather than an artefact of a single vendor's research agenda.

Data leakage and confidentiality breach is the most frequently cited risk category. Approximately 59% of surveyed United States employees admitted to using unapproved AI tools at work, with many sharing work-related information with those tools. Because public AI tools may retain submitted data on external infrastructure with limited contractual assurance over retention, deletion, or training use, this single behaviour pattern creates downstream exposure across intellectual property, regulatory, and security domains simultaneously.

From a technical security perspective, the OWASP Top 10 for LLM Applications identifies ten critical risk categories most relevant to Shadow AI: prompt injection, sensitive information disclosure, supply-chain vulnerabilities, data and model poisoning, improper output handling, excessive agency, system prompt leakage, vector and embedding weaknesses, misinformation, and unbounded consumption. Because Shadow AI tools sit outside the organisation's secure architecture, logging, and monitoring controls by definition, all ten categories are amplified relative to a governed deployment.

Functional coding further shows that Shadow AI manifests differently across business units. In technology and engineering functions, developers increasingly rely on AI coding assistants and debugging tools; the principal risks are source-code leakage, insecure code generation, dependency confusion, and a lack of traceability over AI-generated code - risks illustrated directly by the Samsung source-code exposure episode. Human-resources functions present a distinct risk pattern because AI is frequently used for candidate screening, performance summarisation, and workforce planning, creating exposure to bias, discrimination, and explainability requirements, particularly where AI outputs influence employment decisions.

Legal, compliance, and finance functions report similar concerns regarding privilege exposure, hallucinated legal interpretation, and unauthorised automation of control activities, while marketing and communications functions report risk from copyright infringement, synthetic content misuse, and uncontrolled external messaging. Operations and supply-chain teams face risk from unvalidated demand forecasts and supplier-confidentiality breaches when AI tools are used to analyse third-party data without contractual safeguards. This cross-functional pattern confirms that Shadow AI is not a single-department problem; it is a horizontal operating risk that intersects with every major control function inside the enterprise.

TABLE I

CROSS-FUNCTIONAL SHADOW AI RISK THEMES AND REPRESENTATIVE CONTROLS

Risk Theme	Representative Evidence	Risk Tier	Representative Control
Data leakage & confidentiality breach	59% of US employees admitted using unapproved AI tools at work [7]	High	DLP for prompts/uploads; approved-tool catalogue
Intellectual-property exposure	Samsung source-code exposure episode, 2023 [15]	High	Source-code scanning; contractual data-use terms with approved vendors
Regulatory non-compliance	EU AI Act risk-classification duty [8]	High	AI inventory; risk-tier classification; human oversight assignment
Security vulnerabilities (LLM-specific)	OWASP Top 10 for LLM Applications [9]	High	Red-teaming; output validation; prompt-injection detection
Hallucination / misinformation	Air Canada chatbot liability ruling, 2024 [15]	Medium–High	Human review of all customer-facing and regulatory outputs
Credential & secret exposure	Pasted tokens and API keys into AI prompts [10]	Medium–High	Secrets detection; dynamic vault-based credential issuance
Third-party / supply-chain risk	Compromised apps, APIs, and plug-ins [11]	Medium	Vendor due diligence; API traffic monitoring; provider security review
Operational resilience risk (agentic AI)	Ungoverned agents with write access to enterprise systems [2]	High	Kill switch; transaction limits; human-in-the-loop approval; rollback

C. The Economics of Shadow AI

IBM Security research, corroborated by independent secondary reporting, found that organisations with high Shadow AI exposure incurred breach costs that were, on average, USD 670,000 higher than organisations with little or no Shadow AI exposure, and that one in five surveyed organisations experienced a cyberattack attributable to Shadow AI-related security gaps. The same reporting noted that 97% of organisations that experienced an AI-related breach lacked adequate AI access controls at the time of the incident. Fig. 1b presents this differential graphically using average breach-cost figures derived from the triangulated reporting.

These findings indicate that the economic cost of Shadow AI is not confined to productivity loss or duplicated software spend; it materialises as a measurable increase in the cost of cybersecurity incidents. Shadow AI also introduces hidden costs including remediation, regulatory investigation, legal review, customer notification, model rollback, output correction, incident response, and erosion of stakeholder trust, all of which fall outside a simple subscription-cost comparison.

At the individual level, Shadow AI saves time and increases perceived productivity. At the enterprise level, however, it may create duplicated AI subscriptions, fragmented tooling, inconsistent outputs, hidden data flows, compliance exposure, increased breach costs, technical debt, vendor lock-in, and a lack of audit evidence. This paradoxical economic profile is a key reason why the governance response must balance enablement with control rather than defaulting to prohibition.

D. The Emergent Risk of Autonomous Agents

A consistent theme across the most recent literature is that Shadow AI is evolving from a chatbot problem into an agent problem. Analysis from 2025 describes a progression from consumer-grade AI tools, to business AI gone rogue, to what is being called Shadow Agents: autonomous or semi-autonomous systems built outside IT oversight that can access data, invoke tools, and interact directly with enterprise systems. Unlike a chatbot that responds to a single prompt, an agent can plan a sequence of actions, call external APIs, write and execute code, send communications, and update records, often without a human reviewing each individual step.

This shift requires a new governance primitive. Treating each AI agent as a distinct non-human identity-analogous to a service account or privileged application identity, with its own owner, permissions, logging requirements, and lifecycle - is a necessary condition for control. Without such treatment, an AI agent becomes a shadow identity with unclear ownership, excessive privileges, poor logging, weak authentication, and no lifecycle management. The synthesis identified five practical privilege levels appropriate for agent governance, ranging from Level 0 (text generation only, no system access) to Level 5 (autonomous action in critical systems, requiring executive approval and continuous monitoring), discussed further in Section V.

The OWASP Excessive Agency category specifically highlights the danger of LLM systems being granted too much autonomy or tool access, noting that ungoverned tool permissions are a primary driver of LLM-related security incidents. This is particularly relevant for Shadow Agents, which by definition have not gone through any formal tool-access authorisation process before being allowed to operate within - and act upon - enterprise systems.

E. Regulatory Compliance Gap

Shadow AI intersects with an increasingly active regulatory environment. The EU Artificial Intelligence Act, which entered into force in August 2024, introduced a risk-based framework covering prohibited practices, high-risk systems, transparency obligations, and general-purpose AI model requirements, applicable to deployers as well as providers. Because the Act's obligations depend on accurately classifying each AI system's risk tier, an organisation that lacks a complete AI inventory cannot reliably determine its compliance status - a structural gap directly attributable to Shadow AI.

Under the General Data Protection Regulation (GDPR), individuals have a qualified right not to be subject to decisions based solely on automated processing that produce legal or similarly significant effects, which is directly relevant where Shadow AI tools are used for hiring, performance evaluation, credit, or eligibility decisions. Shadow AI creates GDPR exposure because unauthorised tools may process personal data without a lawful basis, without transparency notices, without data processing agreements, without data minimisation, without retention controls, and without mechanisms for data subject rights exercise.

India's Digital Personal Data Protection Act, 2023, together with its 2025 operating rules, introduces parallel obligations around consent, processor oversight, and breach notification that unmanaged AI tools are structurally unable to satisfy in an auditable manner. ISO/IEC 42001:2023 and the NIST AI Risk Management Framework similarly presuppose the existence of an AI inventory and formal risk-assessment process as foundational governance artefacts - both of which Shadow AI systematically undermines.

Two documented incidents illustrate how regulatory and accountability risk materialises in practice. In 2023, Samsung reportedly restricted employee use of public generative AI tools after engineers uploaded sensitive internal source code into a public chatbot, demonstrating how well-intentioned productivity use can create uncontrolled intellectual-property exposure within days of access being enabled. Separately, in 2024, the British Columbia Civil Resolution Tribunal held Air Canada liable for inaccurate information provided by its website chatbot, rejecting the argument that the chatbot was a legally separate entity and confirming that organisations remain accountable for AI-generated outputs even when those outputs are produced by an automated system. Although the Air Canada matter did not involve unauthorised employee use, it is directly relevant to Shadow AI governance because it demonstrates that accountability for AI output cannot be outsourced to the technology itself.

4. Discussion

The findings support three principal discussion points with direct implications for enterprise practice.

First, the evidence consistently indicates that Shadow AI is driven primarily by a productivity gap rather than employee malfeasance [1], [5]. Employees adopt unsanctioned AI tools because those tools are faster, more capable, or more readily available than the alternatives their organisation provides. This has a direct policy implication: control strategies premised on prohibition alone are likely to fail, because they do not address the underlying demand. The consistent recommendation across synthesised sources is visibility before restriction - discovering existing use through network telemetry, SaaS-discovery tools, and self-disclosure - followed by the rapid provision of an approved, equally capable alternative [2].

Second, the breach-cost data suggests that Shadow AI should be modelled as a quantifiable cyber-risk category rather than treated solely as a policy or cultural issue [11], [12]. The observed breach-cost differential of approximately USD 670,000, combined with the finding that 97% of AI-related breaches involved inadequate access controls, indicates that identity and access management investment specifically targeted at AI tools and agents is likely to deliver a measurable risk-adjusted return. Organisations that have historically justified cybersecurity investment on the basis of breach-probability-times-breach-cost calculations can apply the same framework to Shadow AI governance.

Third, the emergence of autonomous agents represents a discontinuity rather than a simple extension of existing chatbot-era risk [2], [6]. Where a chatbot risk is bounded by what a human chooses to paste into it, an agent risk is bounded by what permissions, tools, and data access the organisation has granted - often implicitly through inherited credentials. This shifts the appropriate control discipline from content-focused data-loss prevention toward identity-focused privileged-access management, a transition that most surveyed organisations have not yet completed. The Air Canada ruling underscores this point: accountability attaches to the deploying organisation regardless of whether the output-producing system is a chatbot, a copilot, or an autonomous agent [15].

Taken together, these findings justify the governance architecture presented in the following section, organised around the principle of converting Shadow AI into Governed AI through staged visibility, risk-tiered enablement, and non-human identity controls for agents - rather than through blanket restriction.

5. Proposed Governance Framework

Based on the thematic synthesis, this section presents a consolidated governance framework comprising ten guiding principles, a risk-tiering structure, and a minimum control baseline for AI agent identity governance.

A. Ten Governing Principles

The framework is organised around ten governing principles derived from the synthesised literature and aligned to recognised external frameworks.

Principle 1: Visibility before restriction: organisations cannot govern AI use they cannot see; discovery should precede any blocking policy. This requires combining network and SaaS telemetry, cloud access security broker signals, endpoint monitoring, and a confidential employee self-disclosure channel.

Principle 2: Enable safe AI rather than merely banning unsafe AI: approved alternatives must be at least as fast and capable as the unsanctioned tools they are intended to replace, otherwise employees will continue to prefer consumer-grade options.

Principle 3: Classify AI use by risk: not all AI use carries equal exposure. Drafting a meeting agenda is low risk. Summarising confidential board material, processing customer data, generating production code, or triggering financial transactions is high risk. A risk-tiered model preserves innovation while constraining high-impact use cases.

Principle 4: Data controls must travel with the prompt: AI governance must integrate with existing data classification schemes so that the system knows whether data is public, internal, confidential, regulated, or privileged, and enforces controls accordingly before submission.

Principle 5: Human accountability remains non-negotiable: AI may generate or recommend, but an accountable human must remain responsible for the resulting business decision, especially in regulated or high-impact contexts.

Principle 6: Treat AI agents as digital identities: agents require their own unique identity, permissions, named owner, expiry date, and access review cycle, rather than inheriting a human user's full privilege set.

Principle 7: Adopt a secure-by-design AI development lifecycle: every AI use case should follow a formal lifecycle covering intake, classification, design review, threat modelling, evaluation, red-teaming, approval, deployment, monitoring, and retirement.

Principle 8: Favour continuous monitoring over one-time approval: models, prompts, data, and integrations change after initial deployment; governance must therefore be ongoing rather than point-in-time.

Principle 9: Embed governance into workflow: approval processes that are slower than the underlying task will be bypassed; controls should appear naturally within the employee's existing tools through approved AI catalogues, prompt templates, and automated risk scoring.

Principle 10: Align to recognised external frameworks: principally the NIST AI Risk Management Framework and its Generative AI Profile, and ISO/IEC 42001 for AI management systems, to provide external and auditable evidence of governance maturity.

B. Risk-Tiering Structure

Table I summarises the eight cross-functional risk themes identified during synthesis, each mapped to an illustrative risk tier and a representative control drawn from the NIST AI RMF control library and OWASP LLM Top 10 mitigations. The tier assignments are indicative; organisations should calibrate tiers to their own regulatory, operational, and sectoral context.

C. Agent Identity Control Baseline

Because autonomous agents represent the most significant emerging risk identified, the framework specifies a minimum control baseline applicable to every enterprise AI agent before it enters production. This baseline comprises six mandatory control groups.

First, each agent must have a unique, enterprise-managed identity. No shared service accounts and no reuse of human user credentials are permitted. The identity record must document the agent name, business owner, technical owner, approved purpose, risk tier, systems accessed, data classification handled, tools permitted, and model used.

Second, credentials must be short-lived, scoped narrowly to the required permissions, issued dynamically through a secrets-management vault, and rotated automatically. No secrets may be embedded in prompts, configuration files, or local scripts.

Third, tool access must be explicitly authorised on a tool-by-tool basis. For each tool, the authorisation record must define allowed and disallowed actions, input and output validation requirements, human approval thresholds, transaction limits, rate limits, logging requirements, rollback procedures, and emergency disablement procedures.

Fourth, human oversight must be calibrated to impact. High-impact actions, external communications, employment decisions, financial transactions, legal positions, security-control changes, require human-in-the-loop approval before execution. Lower-risk automation may operate under human-on-the-loop monitoring, where the agent can act within approved boundaries and a human monitors activity and can intervene.

Fifth, comprehensive, tamper-resistant logs must capture every prompt or instruction, tool invocation, API call, data accessed, decision or recommendation, action executed, approval obtained, output generated, error or exception, and policy check result, with timestamps retained according to legal and audit requirements.

Sixth, every agent with tool access must have a tested kill switch capable of immediate token revocation, API access suspension, tool-access disablement, memory freeze, workflow pause, quarantine of generated outputs, and automatic alert to the owner and security team. The kill switch must be tested prior to production deployment and periodically thereafter.

D. Implementation Roadmap

The synthesised literature converges on a four-phase implementation approach that organisations can adapt to their own baseline maturity.

Phase 1: Discover and Stabilise (0–90 days): launch AI-tool discovery combining network telemetry, SaaS-discovery tooling, expense-report analysis, and employee self-disclosure; issue interim acceptable-use guidance; identify the top twenty AI tools currently in unsanctioned use; block tools presenting clearly unacceptable risk where necessary; establish an approved AI tool list; begin employee awareness communications; stand up a cross-functional AI governance council; and create a confidential AI incident reporting channel.

Phase 2: Govern and Enable (3–6 months): build a formal AI use-case registry with business owner, tool or model, purpose, data types processed, user group, risk classification, regulatory impact, third-party dependencies, human oversight requirement, approval status, and monitoring controls; define the risk-tiering model; implement data-loss prevention controls for AI tools; integrate AI risk assessment into procurement and SaaS review; launch an enterprise-approved generative AI platform; create prompt templates and safe-use playbooks; and begin AI red-teaming for the highest-risk use cases.

Phase 3: Industrialise (6–12 months): automate AI-tool discovery; implement policy-as-code for prompts, data classification, and access; integrate AI governance with existing governance-risk-compliance tooling; create an AI control dashboard for executive reporting; conduct the first internal audit review; implement third-party AI assurance; and monitor model performance, hallucination, drift, and misuse continuously.

Phase 4 : Future-Ready Governance (12–24 months): move toward an ISO/IEC 42001-aligned AI Management System; pursue certification for critical AI workflows; deploy an agent-governance platform; establish an AI model and agent bill of materials; introduce autonomous agent simulation and kill-switch testing as a routine assurance activity; and align AI governance reporting with enterprise resilience, cyber defence, and privacy governance.

6. Conclusions

This paper has examined Shadow AI as a structural enterprise phenomenon arising from a persistent gap between employee productivity demand and organisational AI governance capacity. The synthesis of survey, security, regulatory, and incident-based evidence supports four principal conclusions.

First, Shadow AI is overwhelmingly driven by legitimate productivity pressure rather than malicious intent, and prohibition-only strategies are correspondingly ineffective because they do not address the underlying demand. Employees who cannot access capable, approved AI tools will continue to use their own, and prohibitions that lack an approved alternative tend to drive adoption underground rather than eliminate it.

Second, the risk surface of Shadow AI spans data leakage, intellectual-property exposure, regulatory non-compliance, security vulnerability, and operational resilience simultaneously, and is measurably more expensive when it materialises as a breach. The reported average cost differential of approximately USD 670,000 between organisations with high versus low Shadow AI exposure, combined with the finding that 97% of AI-related breaches involved inadequate access controls, establishes Shadow AI as a quantifiable and addressable cyber-risk category.

Third, the emergence of autonomous AI agents represents a qualitative escalation of the Shadow AI problem, shifting the appropriate control discipline from content-based data-loss prevention toward identity- and privilege-based governance of non-human actors. Agents that inherit broad human credentials, operate without a unique identity, and have no tested kill switch represent a fundamentally different risk class from employees using public chatbots.

Fourth, the regulatory environment : including the EU AI Act, GDPR, India's DPDP Act, ISO/IEC 42001, and the NIST AI Risk Management Framework: increasingly presupposes the existence of an accurate AI inventory and formal risk-assessment process as foundational governance artefacts, both of which Shadow AI systematically undermines.

The overarching implication is that Shadow AI should be treated by enterprise leadership as a structural signal that employees are ready for AI-enabled work faster than most organisations are able to govern it. The appropriate strategic response is visibility, safe enablement, and embedded governance- not fear-based restriction. The future-ready organisation will not ask how to stop employees from using AI; it will ask how to make the safest AI path also the easiest, fastest, and most valuable path.

7. Recommendations

Based on the findings and framework presented, the following eight recommendations are proposed for enterprise leadership, cybersecurity, privacy, and risk functions.

1. Establish an AI discovery capability within the first 90 days, combining network and SaaS telemetry, cloud access security broker signals, expense-report analysis, and a confidential employee self-disclosure channel, to build an accurate baseline of existing Shadow AI use before designing controls.

2. Stand up a cross-functional AI governance council with named representation from cybersecurity, privacy and legal, data governance, procurement, human resources, and internal audit, with explicit authority to approve, restrict, or retire AI use cases.

3. Publish and maintain a living approved AI catalogue that tells employees, in plain language, which tools they may use, for which purposes, and with which categories of data, updated at least quarterly and integrated into employee onboarding.

4. Treat every AI agent as a non-human identity from first deployment, applying the six-group control baseline described in Section V-C, including mandatory kill-switch testing prior to production use.

5. Integrate AI risk assessment into existing third-party and SaaS procurement review so that AI-enabled features embedded inside already-approved software are captured alongside standalone AI products.

6. Require human review and sign-off for any AI-influenced output affecting legal, financial, regulatory, hiring, or customer-facing decisions, consistent with the accountability principle confirmed by the Air Canada tribunal ruling.

7. Align the overall programme to the NIST AI Risk Management Framework and pursue ISO/IEC 42001 alignment for critical AI workflows as a long-term assurance target, to provide external and auditable evidence of governance maturity.

8. Track Shadow AI-related incidents as a distinct category within existing cyber-incident and operational-risk reporting, enabling the organisation to quantify, over time, whether governance investment is reducing both incident frequency and incident cost.

Future research should pursue longitudinal, organisation-level case studies that track the financial and security outcomes of structured Shadow AI governance programmes over multiple years, since the cross-sectional survey evidence synthesised in this paper, while directionally reliable, cannot establish causal effect sizes for any single governance intervention.

Disclaimer: This white paper is based solely on the author's independent research and experience and does not represent Shell's views, positions, or content. The views and opinions expressed in this paper are those of the author and do not necessarily reflect the official position of Shell.

REFERENCES

1. Microsoft and LinkedIn, "2024 Work Trend Index Annual Report: AI at Work Is Here. Now Comes the Hard Part," Microsoft Corp., Redmond, WA, USA, May 2024.
2. Google Cloud, "Shadow AI: From Business Tools Gone Rogue to Shadow Agents," Google Cloud Security Blog, Google LLC, Mountain View, CA, USA, 2025.
3. Gartner, Inc., "Predicts 2025: Navigating Emerging AI Governance and Security Risks," Gartner Res., Stamford, CT, USA, 2024.
4. Infosecurity Magazine, "Cybersecurity Leaders Suspect Shadow AI Use Across Enterprises," Infosecurity Magazine, London, U.K., 2024.
5. KPMG International, "Trust, Attitudes and Use of Artificial Intelligence: A Global Study 2025," KPMG Int'l Ltd., Amstelveen, Netherlands, 2025.
6. McKinsey & Company, "The State of AI in 2025: Agents, Innovation, and Enterprise Transformation," McKinsey Global Survey, McKinsey & Co., New York, NY, USA, 2025.
7. Cybernews, "Shadow AI Survey: How US Employees Use Unapproved AI Tools at Work," Cybernews, Vilnius, Lithuania, 2024.
8. European Commission, "Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (AI Act)," Off. J. Eur. Union, Brussels, Belgium, Aug. 2024.
9. OWASP Foundation, "OWASP Top 10 for Large Language Model Applications, Version 2025," OWASP Found., Wakefield, MA, USA, 2025.
10. Dashlane Inc., "The Shadow AI Risk Report: Credential and Secret Exposure in the Age of GenAI," Dashlane Inc., New York, NY, USA, 2025.
11. IBM Security, "Cost of a Data Breach Report 2025," IBM Corp., Armonk, NY, USA, 2025.
12. Cybersecurity Dive, "Shadow AI Drives Up Breach Costs, IBM Finds," Industry Dive, Washington, DC, USA, 2025.
13. [13]NationalInstitute of Standards and Technology, "Artificial Intelligence Risk Management Framework: Generative AI Profile," NIST AI 600-1, Gaithersburg, MD, USA, 2024.
14. International Organization for Standardization, "ISO/IEC 42001:2023 Information Technology - Artificial Intelligence Management System- Requirements," ISO, Geneva, Switzerland, 2023.
15. Civil Resolution Tribunal of British Columbia, "Moffatt v. Air Canada," 2024 BCCRT 149, Feb. 2024.
16. European Parliament and Council, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Off. J. Eur. Union, Brussels, Belgium, 2016.
17. Government of India, "The Digital Personal Data Protection Act, 2023," Gazette of India, New Delhi, India, 2023.