Chapter 2

# Shadow AI and Organizational Information Security: Risks, Challenges, and Governance Strategies

**Vahid Sinap**[1]

**Abstract**

The rapid diffusion of generative and agentic artificial intelligence has enabled employees to use powerful AI tools outside formal organizational oversight. This phenomenon, known as shadow AI, can improve productivity, creativity, and problem-solving while creating significant risks for information security, privacy, intellectual property, regulatory compliance, and decision quality. This chapter examines the conceptual foundations, organizational drivers, and security implications of shadow AI from a management information systems perspective. It explains how technological accessibility, task–technology misfit, work pressure, inadequate organizational tools, and unclear policies encourage unauthorized AI use. The chapter also discusses risks related to data leakage, unreliable outputs, prompt injection, excessive agency, undocumented integrations, and weak accountability. A risk-based governance approach is proposed, combining clear policies, approved AI tools, technical controls, employee training, human oversight, monitoring, and adaptive authorization mechanisms. The chapter concludes that effective shadow AI management depends on visibility, proportionality, accountability, and employee enablement.

## 1. Introduction

Artificial intelligence (AI) has rapidly evolved from a specialized technological capability into an accessible component of everyday organizational work. Generative AI systems can produce text, images, software code, analyses, and other forms of digital content in response to natural-language instructions, substantially expanding the range of tasks that can be supported by AI (Feuerriegel et al., 2024). Employees now use AI-powered chatbots, coding

1 Assoc. Prof. Dr., Ufuk University, vahidsinap@gmail.com, https://orcid.org/0000-0002-8734-9509

https://doi.org/10.58830/ozgur.pub1366.c5528

assistants, analytical platforms, browser extensions, and productivity applications to prepare reports, summarize documents, analyze data, communicate with customers, and support decision-making. Experimental evidence indicates that generative AI can reduce the time required to complete professional writing tasks while improving output quality (Noy & Zhang, 2023). Similarly, a large-scale study of customer-support employees found that access to a generative AI assistant increased worker productivity, although the magnitude of this improvement differed across employees (Brynjolfsson et al., 2025). These capabilities make AI tools attractive not only to organizations pursuing formal digital transformation initiatives but also to individual employees seeking faster and more effective ways to perform their work.

The accessibility of these technologies has gradually shifted part of organizational AI adoption away from centrally coordinated information technology processes. Employees can begin using publicly available AI applications, personal subscriptions, external application programming interfaces, and AI functions embedded in third-party software without requiring substantial technical expertise or organizational infrastructure. As a result, the AI tools formally approved by an organization may differ considerably from those actually used in its daily operations. The use of AI tools, models, or applications without the knowledge, authorization, or oversight of relevant organizational units is commonly described as shadow AI (Puthal et al., 2025). These units may include information technology, cybersecurity, legal, data governance, procurement, and regulatory compliance departments.

Shadow AI is conceptually rooted in the broader phenomenon of shadow information technology. Shadow IT refers to technological systems, applications, or services that employees use or develop outside formally authorized organizational IT arrangements (Haag & Eckhardt, 2017). However, shadow AI extends this phenomenon by introducing systems that do more than store, transfer, or present information. AI applications can process organizational data, identify patterns, generate new content, recommend actions, and influence human decisions. Consequently, the risks associated with shadow AI may continue to affect an organization even after the initial interaction with an unauthorized tool has ended. Data submitted to an external AI service may be retained or processed beyond the organization's direct control, while AI-generated outputs may subsequently be incorporated into reports, software, communications, or decision processes without a clear record of their origin.

The emergence of shadow AI should not be explained exclusively through employee negligence or deliberate noncompliance. Research on shadow IT shows that employees frequently adopt unauthorized technologies because formally provided systems do not adequately meet their operational needs (Haag & Eckhardt, 2024). Employees may also encounter lengthy approval procedures, limited access to organizational AI systems, insufficient technical support, or tools that do not offer the functionality required for a particular task. At the same time, demands for higher productivity, faster task completion, experimentation, and innovation may encourage employees to adopt immediately available AI applications. Unauthorized technology use can therefore reflect an attempt to resolve a task–technology mismatch rather than an intention to harm the organization. Haag and Eckhardt (2024) consequently argue that shadow technology should be managed by addressing both cybersecurity requirements and legitimate user needs instead of relying solely on restrictive controls.

This perspective reveals a fundamental tension surrounding shadow AI. On the one hand, employee-driven AI adoption may facilitate experimentation, creativity, learning, and rapid problem-solving. It can also reveal unmet technological needs that formal organizational systems have failed to address. On the other hand, the absence of organizational oversight may expose sensitive information, intellectual property, personal data, internal communications, source code, customer records, and strategic documents to external providers. Puthal et al. (2025) associate shadow AI with data leakage, security breaches, regulatory noncompliance, model vulnerabilities, and an expanded organizational attack surface. The risks are not limited to the confidentiality of data. Generative AI systems may also produce inaccurate, biased, fabricated, or misleading outputs, creating threats to information integrity and the reliability of organizational decisions. The Generative Artificial Intelligence Profile published by the National Institute of Standards and Technology identifies risks concerning data privacy, information security, confabulation, intellectual property, harmful bias, and human overreliance as important considerations in the organizational use of generative AI (National Institute of Standards and Technology [NIST], 2024).

Limited visibility makes these risks particularly difficult to manage. When an AI application is adopted outside formal organizational processes, managers may be unable to determine which tools are being used, what information is entered into them, where the information is processed, how long it is retained, or whether generated outputs are verified before use. Unauthorized applications may also bypass existing controls for vendor assessment, access authorization, data classification, procurement, incident reporting, and

regulatory compliance. Traditional information security arrangements generally protect known systems, recognized users, and observable data flows. Shadow AI creates an oversight gap because the organization cannot effectively assess or control AI uses whose existence it has not identified. This gap becomes increasingly important as AI tools are integrated into routine workflows and begin to affect organizational knowledge, business processes, and managerial decisions.

Shadow AI should therefore be viewed as more than an isolated cybersecurity problem. From a management information systems perspective, it constitutes a sociotechnical challenge involving the interaction of employees, organizational structures, work requirements, data, digital technologies, and governance mechanisms. Managing this challenge requires technical controls, but technical restrictions alone are unlikely to eliminate informal AI use. Organizations must also understand why employees adopt unauthorized tools, distinguish low-risk experimentation from high-risk practices, provide secure alternatives, establish clear responsibilities, and create accessible procedures for approving new AI applications. The NIST AI Risk Management Framework emphasizes that effective AI risk management should be integrated into organizational policies and processes through the interconnected functions of governing, mapping, measuring, and managing AI risks (NIST, 2023). Its generative AI profile further stresses that risk-management practices must be adapted to the context, objectives, legal obligations, and risk tolerance of each organization (NIST, 2024).

Against this background, this chapter examines shadow AI as an emerging organizational information security phenomenon. It first clarifies the concept and explains the organizational conditions that encourage employees to use unauthorized AI tools. It then evaluates the implications of shadow AI for information confidentiality, privacy, intellectual property, regulatory compliance, cybersecurity, and decision integrity. Finally, it discusses governance strategies through which organizations can increase the visibility and accountability of AI use without unnecessarily suppressing employee initiative and digital innovation. Rather than assuming that all informal AI use is inherently harmful, the chapter adopts a balanced perspective that recognizes both the operational value and the security consequences of employee-driven AI adoption. In doing so, it positions shadow AI at the intersection of information systems management, employee technology behavior, organizational innovation, cybersecurity, and AI governance.

## 2. Conceptual Foundations and Organizational Drivers of Shadow AI

### 2.1. Conceptual Boundaries and Manifestations of Shadow AI

Shadow AI can be understood as a contemporary extension of shadow information technology, but the two concepts should not be treated as completely interchangeable. Shadow IT broadly refers to software, hardware, or digital services that organizational members acquire, develop, or use without alignment with the formal IT function (Klotz et al., 2019). Shadow AI represents a more specific form of this phenomenon in which the unauthorized or undisclosed technological resource possesses AI-based capabilities such as content generation, prediction, classification, recommendation, decision support, or autonomous task execution (Puthal et al., 2025). The defining characteristic is therefore not simply the presence of AI but the absence of appropriate organizational visibility, authorization, or governance.

This distinction is important because the use of an externally developed or employee-selected AI application does not automatically constitute shadow AI. An employee may identify an AI tool independently and subsequently disclose it to the relevant organizational units, obtain approval, and use it under agreed security and data-management conditions. Such an arrangement is more appropriately regarded as employee-initiated or business-managed AI rather than shadow AI. Klotz et al. (2019) similarly distinguish covert shadow IT from business-managed IT, in which business units assume responsibility for technological resources while remaining aligned with the formal IT organization. Applying this distinction to AI suggests that organizational alignment, rather than the original source of the technology, determines whether an application remains in the shadows.

Conversely, an officially available AI system can generate shadow AI practices when it is used beyond its approved purpose or under conditions that evade organizational controls. An employee may, for example, use an approved generative AI assistant but enter a category of data that organizational policy prohibits, connect the system to an unauthorized external source, or rely on its output for a decision for which human review is mandatory. Shadow AI may therefore involve both the adoption of an unapproved technology and the unapproved use of an otherwise authorized technology. This broader interpretation is consistent with shadow IT taxonomies that include not only unofficial systems but also the misuse or unintended use of official technological resources (Klotz et al., 2019).

The boundary of shadow AI can be clarified through four interrelated criteria. The first is authorization, referring to whether the application, model, or use case has received formal approval. The second is visibility, which concerns whether IT, cybersecurity, data-governance, or managerial units know that the AI system is being used. The third is governance alignment, referring to whether the use is subject to organizational requirements concerning data handling, vendor assessment, access control, accountability, and human oversight. The fourth is purpose alignment, which indicates whether the AI system is being used for an approved organizational task and within its authorized scope. A practice may therefore become shadow AI when one or more of these conditions are absent, even when the employee does not deliberately intend to conceal the technology.

Shadow AI may appear in multiple forms, ranging from occasional interactions with public chatbots to AI-enabled workflows embedded in routine business processes. Some uses are direct and visible to the employee, such as entering a document into a public generative AI service. Others are less apparent because AI capabilities are integrated into browser extensions, office applications, analytical platforms, customer-management systems, or software-development tools. The growing modularity of AI also allows employees to connect external models to organizational data through application programming interfaces, no-code platforms, custom assistants, and automated agents. Puthal et al. (2025) emphasize that shadow AI can include unauthorized models, tools, and systems operating beyond the supervision of centralized IT and cybersecurity functions.

***Table 1 Common Manifestations of Shadow AI in Organizations***

| **Manifestation** | **Illustrative employee practice** | **Why the practice constitutes shadow AI** |
|---|---|---|
| Public generative AI services | Using a personal account to summarize internal reports, generate correspondence, or analyze organizational documents | The service and its data-processing conditions have not been evaluated or approved by the organization |
| AI-enabled browser extensions and add-ons | Installing an extension that reads webpages, emails, documents, or meeting content to generate summaries or responses | The AI capability may access organizational information without appearing as a separate organizational system |
| External coding and analytical tools | Uploading source code, datasets, or technical logs to an external AI assistant for debugging or analysis | The tool operates outside approved development, data-analysis, and vendor-management environments |

| | | |
|---|---|---|
| Employee-built assistants and models | Creating a custom chatbot, fine-tuned model, or retrieval-based assistant using organizational documents | The resulting system may process or reproduce organizational knowledge without registration, testing, or formal ownership |
| AI-supported no-code automations | Connecting an external model to email, cloud storage, customer records, or business applications through a no-code platform | The integration creates an unapproved data flow and may automate actions outside existing controls |
| Autonomous or semi-autonomous AI agents | Configuring an agent to retrieve information, communicate with third parties, modify records, or perform multistep tasks | The agent may exercise delegated authority without formal approval, monitoring, or clearly assigned accountability |

***Note. Developed by the author based on Klotz et al. (2019), Puthal et al. (2025), and Waters-Lynch et al. (2025). The categories are not mutually exclusive, as a single shadow AI practice may combine several tools and forms of automation.***

Shadow AI may also differ in its scale, duration, and degree of autonomy. It can be limited to a single employee completing a one-time task or become a shared practice adopted by an entire team. Similarly, some practices remain temporary experiments, whereas others become embedded in recurring workflows and gradually assume operational importance. The latter situation is particularly significant because an AI tool may begin as an informal productivity aid but eventually become an undocumented dependency for a business process. At that point, discontinuing or controlling the tool may become difficult because employees, information flows, and operational routines have already adapted to its presence.

The concept of shadow user innovation further illustrates that informal AI use may extend beyond simple technology adoption. Waters-Lynch et al. (2025) define shadow user innovation as covert, employee-initiated value creation enabled by digital tools that can be used and concealed relatively easily. From this perspective, employees do not merely select an unauthorized application; they may design new workflows, combine models with organizational knowledge, and develop task-specific AI solutions. Such activities can reveal valuable opportunities for organizational learning and innovation. Nevertheless, their covert character prevents the organization from evaluating, documenting, scaling, or governing the resulting practices. Shadow AI should consequently be understood as a continuum that ranges from informal individual assistance to the concealed redesign of organizational processes.

### 2.2. Organizational and Behavioral Drivers of Shadow AI

The emergence of shadow AI cannot be adequately explained by a single motive. A useful starting point is the framework developed by Klotz et al. (2019), which categorizes the causes of shadow IT into enablers, motivators, and missing barriers. Enablers make unauthorized technology adoption technically or practically possible; motivators create reasons for employees to adopt it; and missing barriers reduce the likelihood that the behavior will be prevented or redirected. This framework is particularly suitable for shadow AI because its growth reflects a combination of technological accessibility, work-related benefits, and gaps in organizational governance.

The primary technological enabler is the increasing accessibility of powerful AI capabilities. Generative AI applications can be accessed through ordinary web browsers, mobile applications, personal accounts, browser extensions, and embedded software features. Natural-language interfaces further reduce the expertise needed to use these systems. Employees no longer need to develop a model or request extensive technical support to perform activities such as summarization, translation, coding, content generation, or data interpretation. As Klotz et al. (2019) observe in the broader shadow IT context, declining technological complexity and the expansion of easily accessible digital services enable business users to deploy technological solutions independently. AI intensifies this trend by transforming natural language into an interface for technological development and automation.

The concealability of generative AI constitutes another important enabler. Unlike a conventional unauthorized information system, which may require software installation, dedicated infrastructure, or a visible organizational project, many AI tools can be used through a personal browser session or an existing software platform. Waters-Lynch et al. (2025) argue that generative AI enables covert employee innovation partly because these tools are readily accessible and relatively easy to conceal. Furthermore, AI functions may be embedded within applications already used by the organization, making it difficult for employees and managers to recognize when an ordinary digital tool has begun processing organizational information through an external AI model.

Although accessibility makes shadow AI possible, employees generally require a work-related motivation to use it. Performance expectancy is particularly important. Employees may believe that an AI application will allow them to complete tasks more rapidly, improve the quality of their outputs, overcome skill limitations, or manage an excessive workload. Nguyen (2024) found that performance expectancy and effort expectancy significantly

influenced employees' intentions to use shadow IT. These findings are highly relevant to shadow AI because generative AI combines potentially high performance benefits with relatively low usage effort. The perceived balance between substantial task-related value and minimal adoption cost can make informal AI use especially attractive.

A mismatch between employee needs and formally available technologies may strengthen this motivation. Official systems may lack the required functionality, be difficult to use, or respond too slowly to emerging task requirements. Approval and procurement procedures may also be incompatible with the speed at which employees are expected to deliver results. Haag and Eckhardt (2024) show that shadow IT frequently arises when employees attempt to overcome work-related challenges that authorized technologies do not adequately address. Accordingly, the continued use of shadow AI may signal not only a security problem but also weaknesses in the organization's technological support, user experience, or responsiveness to business needs.

Innovation-related motives also deserve attention. Employees may use AI to test ideas, create prototypes, explore alternative solutions, or develop new methods without waiting for formal authorization. In such cases, shadow AI enables local experimentation and may help employees respond to opportunities that centralized IT structures have not yet recognized. Waters-Lynch et al. (2025) conceptualize this behavior as a form of user innovation capable of contributing to organizational capability renewal. However, the same employees may choose to conceal their experiments because they expect managerial resistance, fear that the activity will be prohibited, or believe that formal disclosure will introduce delays. Innovation-oriented use and policy avoidance can therefore occur simultaneously.

Social influences may further normalize shadow AI. When colleagues regularly use generative AI tools, employees may interpret this behavior as acceptable even when no formal policy permits it. Nguyen (2024) found that subjective norms significantly affected shadow IT usage intention, suggesting that employee behavior is influenced by perceptions of what relevant others expect or consider normal. Informal encouragement from supervisors may have a similar effect. A manager who prioritizes rapid results while remaining silent about how those results are achieved may unintentionally communicate that unauthorized AI use is tolerable. The absence of visible negative outcomes can reinforce this perception and reduce the employee's sense that formal approval is necessary.

Employees may also view their behavior as reasonable because they do not intend to damage the organization. Barlette et al. (2025) describe shadow

IT as the voluntary adoption of unapproved tools for greater efficiency, even though such adoption may violate organizational security policies. Their findings indicate that users can perceive shadow technology as both beneficial and threatening and may attempt to reduce some risks while preserving the efficiency advantages they obtain. This suggests that employees are not always indifferent to information security. Instead, they may rely on their own informal risk assessments, such as removing obvious identifiers from a document or avoiding certain categories of information. The problem is that these individual precautions may not reflect the actual technical, contractual, or regulatory conditions under which the AI provider processes data.

Missing organizational barriers form the third group of drivers. Shadow AI is more likely to emerge when organizations have no clear AI-use policy, when employees do not understand existing rules, or when responsibilities are fragmented among IT, cybersecurity, legal, procurement, and data-governance units. A general instruction to "use AI responsibly" may be insufficient because it does not explain which tools are authorized, what data may be entered, which outputs require verification, or how employees can request approval for a new use case. Weak monitoring and limited inventories of AI applications also allow unauthorized practices to continue without detection.

Paradoxically, highly restrictive policies may not eliminate shadow AI when they are unaccompanied by usable alternatives. Haag and Eckhardt (2024) conclude that shadow IT cannot be addressed through a universal control strategy and recommend balancing cybersecurity requirements with employee needs. When employees perceive that compliance prevents them from completing legitimate tasks, restrictions may encourage concealment rather than secure behavior. Effective governance must therefore reduce the organizational conditions that make shadow AI necessary or attractive. This requires accessible approved tools, proportionate approval procedures, clear data-use boundaries, responsive technical support, and opportunities for employees to disclose useful AI experiments without automatically facing sanctions.

Taken together, these drivers show that shadow AI is produced by the interaction of technological opportunity, individual expectations, social norms, work pressures, and organizational limitations. It should not be reduced to either employee misconduct or technological inevitability. Shadow AI becomes more likely when highly accessible AI tools offer immediate task-related value while formal organizational arrangements remain slow, unclear, or poorly aligned with users' needs. Understanding this combination is essential

because governance strategies that target only employee behavior will leave the underlying technological and organizational causes unchanged.

## 3. Organizational Information Security Risks

Shadow AI changes the nature of organizational information security risk because it creates data flows, technological dependencies, and AI-supported decisions that remain partly or entirely outside formal oversight. Conventional cybersecurity practices are generally designed around identifiable assets, registered users, approved vendors, and documented information flows. In shadow AI environments, however, the organization may not know which applications are being used, what information is being transferred, what external services are processing that information, or how AI-generated outputs are entering organizational processes. The central problem is therefore not only that an AI system may be technically vulnerable. It is also that the organization cannot effectively assess, monitor, or respond to a system whose use has not been formally identified.

These risks affect the three conventional objectives of information security: confidentiality, integrity, and availability. Shadow AI may compromise confidentiality when employees disclose organizational or personal data to unauthorized services. It may weaken integrity when inaccurate, manipulated, or unverified AI outputs are incorporated into organizational records and decisions. It may also affect availability and operational continuity when business processes become dependent on external AI applications that the organization does not control. In addition, generative AI introduces risks relating to privacy, intellectual property, human oversight, regulatory accountability, and autonomous action that extend beyond traditional cybersecurity boundaries (National Institute of Standards and Technology [NIST], 2024; Puthal et al., 2025).

### 3.1. Confidentiality, Privacy, and Intellectual Property Exposure

The most immediate risk associated with shadow AI arises when employees submit organizational information to an unauthorized external service. Prompts may contain customer records, employee information, meeting transcripts, financial figures, contractual documents, software code, product designs, internal policies, strategic plans, or research findings. Even when the employee uses the information only to request a summary, translation, analysis, or rewritten text, the interaction constitutes a transfer of information to a third-party technological environment. Because the tool has not been formally assessed, the organization may lack reliable information about where the data are processed, how long they are retained, whether they are used for service

improvement or model training, and which subcontractors or jurisdictions are involved.

This uncertainty distinguishes shadow AI from approved enterprise AI services. Formal procurement and security-assessment processes can evaluate contractual terms, retention periods, access controls, encryption, incident-notification procedures, and restrictions on secondary data use. Shadow AI bypasses these processes. As a result, the organization may be unable to determine whether an external provider's practices are consistent with its data-classification policies, contractual commitments, or legal obligations. The European Data Protection Board's technical report on large language models emphasizes that privacy risks must be assessed by examining data flows across the entire system lifecycle and identifying the parties that process, store, or receive the information (Barberá, 2025). Such an assessment becomes difficult when employees introduce AI services without disclosure.

Personal data deserve particular attention. Under the General Data Protection Regulation, personal data must be processed lawfully, fairly, and transparently and must be collected for specified purposes while remaining adequate, relevant, and limited to what is necessary (European Parliament & Council of the European Union, 2016). When employees enter personal information into an unauthorized AI system, the organization may be unable to establish the lawful basis for this additional processing or determine whether the use is compatible with the original purpose for which the data were collected. It may also be unable to provide accurate information to affected individuals, respond to data-subject requests, establish appropriate processing agreements, or verify international data-transfer conditions.

The problem is not limited to obviously identifiable records. Employees may assume that removing a person's name is sufficient to make a document safe for external AI processing. However, contextual details, job titles, transaction histories, locations, unusual events, and combinations of indirect identifiers may permit reidentification. AI systems can also generate inferences about individuals from seemingly ordinary information. Barberá (2025) therefore treats excessive data collection, insufficient anonymization, unauthorized access, lack of transparency, and the exposure of sensitive attributes as distinct but interconnected privacy risks. Informal employee judgments about whether information is "anonymous enough" may not provide adequate protection.

Confidentiality risks can also emerge after data have entered an AI system. Large language models may reproduce or reveal information contained in their training or adaptation data under particular conditions. Carlini et al. (2021) demonstrated that individual training examples, including personally

identifiable information and source code, could be extracted from a language model through carefully designed queries. This finding does not mean that every prompt submitted to every commercial AI service will later be disclosed. It does demonstrate, however, that model memorization and data extraction constitute technically plausible risks and that organizations should not assume that data become irretrievable merely because they have been processed by a model.

Sensitive information may also be exposed through AI-generated outputs. The OWASP Foundation (2024) identifies sensitive information disclosure as a major vulnerability of applications based on large language models. Such disclosure may involve personal data, financial information, health records, legal documents, security credentials, proprietary algorithms, or confidential business information. In a shadow AI context, the risk is intensified because the organization may not have configured output restrictions, access controls, data filters, or logging mechanisms. An employee-built assistant connected to a shared document repository, for example, may return information to users who would not have been authorized to access the original files.

Intellectual property risks overlap with confidentiality risks but are not identical to them. Employees may upload copyrighted materials, unpublished manuscripts, product specifications, proprietary datasets, software code, formulas, designs, or trade secrets to generate new content or obtain technical assistance. NIST (2024) notes that generative AI may create intellectual property risks through the use of protected material in system inputs, training processes, and generated outputs. An organization may therefore face uncertainty about whether it has the right to submit particular material to the service, whether generated content reproduces protected material, and who owns or may reuse the resulting output.

The exposure of trade secrets is especially significant because the economic value of such information depends on its continued secrecy and controlled use. An employee may believe that submitting a limited excerpt of code or a partial commercial strategy presents little risk. Yet repeated interactions by multiple employees can collectively reveal substantial elements of an organization's knowledge base. The absence of centralized visibility prevents the organization from understanding this cumulative exposure. Shadow AI may consequently produce a gradual form of information leakage in which no single prompt appears catastrophic, but the aggregated flow of prompts, uploaded documents, custom instructions, and model integrations reveals valuable organizational knowledge.

Confidential communications may also be affected. Employees in human resources, legal, finance, healthcare, education, or management positions may use public AI systems to summarize disputes, draft responses, evaluate cases, or prepare recommendations. These interactions may disclose information protected by professional, contractual, or sector-specific confidentiality requirements. Moreover, the organization may be unable to preserve an appropriate audit trail showing what information was disclosed and how it was processed. The informational value of an AI interaction should therefore be assessed not only by examining the individual prompt but also by considering its context, the identity of the affected parties, and the sensitivity of the organizational process in which it occurs.

### 3.2. Information Integrity, Decision Reliability, and Regulatory Accountability

Shadow AI creates a second category of risk by allowing AI-generated content to enter organizational workflows without systematic verification or provenance records. Generative AI systems produce outputs by estimating statistically plausible continuations rather than by independently confirming the truth of every statement. As a result, they may generate inaccurate facts, fabricated references, incorrect calculations, false explanations, or internally inconsistent recommendations. NIST (2024) uses the term confabulation to describe confidently presented but erroneous or false generative AI content. Farquhar et al. (2024) similarly demonstrate that large language models can generate plausible but arbitrary and incorrect answers, particularly when reliable information is unavailable.

The organizational consequence depends on how the output is used. An inaccurate sentence in an informal brainstorming exercise may have limited impact, whereas an inaccurate output incorporated into a financial report, legal document, customer communication, software application, employee assessment, or strategic recommendation may produce significant harm. Shadow AI makes this distinction difficult to manage because the same public tool can be used for both low-impact and high-impact tasks without a formal change in authorization. Employees may initially adopt the tool for editing text and later begin relying on it for interpretation, analysis, or recommendations.

The fluent and confident presentation of AI-generated content may also encourage automation bias. Users may interpret detailed, well-structured, and professionally worded responses as evidence of accuracy even when the system provides no verifiable basis for its claims. NIST (2024) observes that users may over-rely on generative AI and perceive its outputs as being of

higher quality than they actually are. This tendency is especially important when employees operate under time pressure or lack the domain expertise required to identify subtle errors. Shadow AI may therefore substitute apparent efficiency for careful verification.

Errors may also become difficult to trace after AI-generated material has been edited or integrated into other documents. An employee may copy part of an AI response into a report, revise the wording, and remove any indication that AI was involved. The resulting document appears to be an ordinary organizational output even though some of its claims originated in an external probabilistic system. When an error is later discovered, managers may be unable to reconstruct the prompt, the model version, the data sources, or the reasoning that produced it. This loss of provenance undermines accountability and prevents the organization from learning systematically from AI-related incidents.

The integrity problem is not limited to factual errors. AI-generated outputs can contain biased assumptions, inappropriate generalizations, insecure code, or recommendations that are unsuitable for the organizational context. A model may generate a technically plausible policy that conflicts with internal procedures, summarize a contract while omitting a critical exception, or produce software code containing exploitable weaknesses. Because the tool is unauthorized, these outputs may not be subjected to the testing, validation, and human oversight requirements that would apply to an approved organizational system.

Information integrity may also be deliberately attacked. External content used by an AI application may contain hidden or manipulative instructions designed to alter the system's behavior. An employee may ask an AI assistant to summarize a webpage, document, email, or shared file without realizing that the content includes instructions intended for the model rather than the human reader. Greshake et al. (2023) show that this blurring of the boundary between data and instructions creates indirect prompt-injection vulnerabilities in applications integrated with large language models. A malicious document can therefore influence the AI-generated summary, redirect the system's behavior, or cause it to disclose information from connected sources.

Regulatory accountability becomes more difficult when the organization cannot document how AI is being used. The EU AI Act adopts a risk-based approach and assigns different responsibilities according to the nature of the AI system and the role of the organization using it (European Parliament & Council of the European Union, 2024). Shadow AI can prevent an organization from determining whether a particular use falls within a regulated

category or whether obligations concerning documentation, human oversight, transparency, monitoring, or risk management apply. Even when the AI Act does not classify the use as high-risk, other legal frameworks may continue to govern personal data, employment practices, consumer protection, intellectual property, professional responsibility, and sectoral confidentiality.

The governance gap is particularly important when AI outputs influence decisions about individuals. Employees may informally use AI to screen applications, summarize performance records, evaluate customer complaints, interpret medical or educational information, or recommend personnel actions. The final decision may still be made by a human, but the AI system may shape which information is emphasized and which alternatives are considered. If this use remains undisclosed, the organization may be unable to evaluate fairness, accuracy, explainability, or the adequacy of human oversight. It may also be difficult to respond when an affected individual asks how the decision was reached.

Shadow AI may thus create a form of accountability fragmentation. The employee selects the tool, an external provider operates the model, organizational data supply the context, and a manager may rely on the result. Yet no party within the organization has formally accepted responsibility for evaluating the entire process. When harm occurs, uncertainty may arise concerning whether responsibility belongs to the employee, the manager, the IT department, the data owner, the organization, or the AI provider. This fragmentation is not simply a legal issue; it is an information systems problem because effective accountability requires clearly defined ownership of data, systems, outputs, and decisions.

### 3.3. Expanded Attack Surface and Operational Exposure

Shadow AI also expands the organizational attack surface by creating connections between external AI systems and internal information resources. A simple chatbot interaction may involve only text entered manually by an employee. More advanced shadow AI practices may connect a model to email, cloud storage, source-code repositories, customer databases, calendars, web browsers, or enterprise applications. Each integration creates additional pathways through which data can be accessed, instructions can be manipulated, and actions can be executed. Because these integrations are not formally registered, security teams may not include them in asset inventories, threat models, penetration tests, access reviews, or incident-response plans.

Prompt injection is one of the most prominent risks in this environment. A direct prompt injection occurs when a user deliberately supplies instructions

intended to override the model's expected behavior. An indirect prompt injection occurs when the malicious instruction is embedded in external content later retrieved by the model (OWASP Foundation, 2024). The latter is particularly relevant to employee-built assistants and no-code automations because the employee may trust the AI system to read websites, emails, or documents from multiple sources. A successful injection can manipulate outputs, disclose sensitive information, access unauthorized functions, influence decisions, or initiate commands in connected systems.

Retrieval-augmented generation does not eliminate this problem. Although connecting an AI system to an internal knowledge base can improve the relevance of its responses, it also creates a pathway through which malicious or improperly classified content can affect the generated output. If access controls are poorly implemented, the assistant may retrieve information beyond the requesting employee's authorization. If documents within the knowledge base contain hidden instructions, the model may interpret them as commands. Greshake et al. (2023) demonstrate that application-integrated language models can be remotely manipulated through content that the model retrieves and processes.

The growing use of autonomous and semi-autonomous AI agents further increases operational risk. An AI agent may be permitted to call external tools, send messages, access records, create files, modify databases, or complete multistep tasks. The OWASP Foundation (2024) describes excessive agency as a vulnerability arising when an AI system has more functionality, permissions, or autonomy than is necessary. When an unauthorized agent operates through an employee account, the organization may not know that decision authority has effectively been delegated to an external model. A hallucinated instruction, manipulated input, or compromised plugin can then produce actions rather than merely inaccurate text.

The consequences of excessive agency extend across confidentiality, integrity, and availability. An agent with broad permissions may read confidential documents, modify records, send unauthorized communications, approve transactions, or delete files. The risk is particularly severe when the agent uses a shared or privileged identity rather than acting within the authorization scope of the individual employee. Human confirmation may also be absent if the employee has configured the workflow to maximize speed. A shadow AI assistant can therefore evolve from a personal productivity tool into an unmonitored actor within the organization's digital environment.

Supply-chain vulnerabilities constitute another source of exposure. AI applications frequently depend on external models, plugins, browser extensions,

open-source libraries, datasets, vector databases, and application programming interfaces. The employee may evaluate the visible functionality of the tool without understanding these underlying dependencies. A compromised extension, insecure software component, manipulated model, or change in the provider's service may affect organizational data and workflows. The organization may have no contractual right to receive notice of such changes because the service was adopted through a personal or free account.

Data and model poisoning can similarly undermine integrity. An attacker may manipulate training, fine-tuning, retrieval, or embedding data to introduce biased behavior, hidden triggers, or misleading outputs. NIST (2024) identifies data poisoning as a threat capable of altering a generative AI system's operation, while the OWASP Foundation (2024) includes data and model poisoning among the principal security risks for large language model applications. Shadow AI increases this risk because employees may construct custom assistants using datasets whose quality, provenance, and security have not been examined.

Improper handling of AI-generated outputs can create downstream vulnerabilities even when the model itself has not been directly compromised. Generated code may be executed without security testing; model-produced queries may be passed to databases; generated HTML may be displayed in an application; and AI-generated instructions may be sent to other automated systems. When outputs are treated as trusted rather than untrusted content, they can enable code execution, unauthorized requests, or data corruption. This risk becomes greater in no-code and low-code environments where employees can connect multiple services without understanding the security implications of each connection.

Availability and business continuity also require consideration. Employees may gradually embed unauthorized AI tools into essential workflows, making task completion dependent on services that the organization does not manage. The provider may change prices, functionality, usage limits, model behavior, or access conditions without organizational planning. Accounts may be suspended, services may become unavailable, or employees who created the workflow may leave the organization without documenting it. The organization may then discover that an important process depends on a personal subscription, undocumented prompt library, or external automation for which no alternative exists.

Shadow AI incidents may also be difficult to detect and investigate. Traditional monitoring systems may record that an employee visited an AI website but may not reveal the meaning or sensitivity of the information

entered into a prompt. Personal accounts and devices further reduce visibility. When an incident occurs, security teams may lack logs showing which data were disclosed, which model processed them, what outputs were generated, and whether those outputs were shared or acted upon. This weakens containment, notification, evidence preservation, and post-incident analysis.

***Table 2 Principal Organizational Information Security Risks Associated with Shadow AI***

| **Risk domain** | **Shadow AI mechanism** | **Organizational assets affected** | **Potential consequences** |
|---|---|---|---|
| Confidential data disclosure | Employees enter internal documents, records, code, or strategic information into unauthorized AI services | Commercially sensitive information, customer data, source code, internal communications | Data leakage, loss of confidentiality, contractual breaches, reputational damage |
| Privacy and personal data misuse | Personal data are processed without an established purpose, lawful basis, transparency process, or approved provider relationship | Customer, employee, patient, student, or applicant information | Privacy violations, inability to fulfil data-subject rights, regulatory exposure |
| Intellectual property exposure | Protected content, trade secrets, designs, datasets, or unpublished materials are submitted to external systems or reproduced in outputs | Copyrighted works, patents, trade secrets, proprietary knowledge | Loss of control over intellectual assets, infringement claims, weakening of competitive advantage |
| Information integrity failure | Confabulated, biased, incomplete, or contextually inappropriate outputs enter reports and decisions | Organizational records, analyses, policies, software, managerial decisions | Incorrect decisions, operational errors, unreliable records, stakeholder harm |
| Loss of provenance and accountability | AI involvement, prompts, data sources, model versions, and human review are undocumented | Audit trails, decision records, governance responsibilities | Inability to explain or reproduce decisions, fragmented responsibility, weak incident learning |
| Prompt injection and manipulation | Malicious instructions are entered directly or embedded in websites, emails, and documents retrieved by the AI system | Connected data sources, model behavior, downstream applications | Unauthorized disclosure, manipulated output, control bypass, command execution |

| | | | |
|---|---|---|---|
| Excessive agency | Unauthorized agents receive broad permissions to access systems or execute actions | Email, databases, cloud storage, business applications, user accounts | Unauthorized transactions, data modification or deletion, operational disruption |
| AI supply-chain exposure | Employees rely on unassessed models, plugins, extensions, APIs, libraries, or datasets | AI workflows, credentials, internal systems, processed data | Compromise through third parties, malicious updates, hidden dependencies |
| Data and model poisoning | Manipulated training, retrieval, fine-tuning, or embedding data alter model behavior | Knowledge bases, custom assistants, analytical outputs | Persistent misinformation, biased results, hidden backdoors, loss of model reliability |
| Availability and continuity risk | Critical work becomes dependent on personal accounts or uncontrolled external services | Business processes, employee knowledge, organizational productivity | Service interruption, vendor lock-in, undocumented dependencies, loss of operational capability |
| Incident-response blind spots | AI use and prompt content are not visible in formal logs or asset inventories | Security monitoring, evidence, incident records | Delayed detection, incomplete containment, inaccurate breach assessment |

***Note. Developed by the author based on Barberá (2025), Greshake et al. (2023), NIST (2024), OWASP Foundation (2024), and Puthal et al. (2025).***

The risks summarized in Table 2 are interdependent rather than isolated. A single shadow AI practice can simultaneously expose confidential data, violate privacy requirements, introduce inaccurate information, and create a new attack pathway. For example, an employee-built agent connected to a customer database may disclose personal information through an external model, generate an incorrect response, and perform an unauthorized action after receiving a manipulated instruction. The severity of shadow AI therefore depends not only on the selected tool but also on the sensitivity of the data, the employee's permissions, the degree of system integration, the autonomy granted to the AI, and the importance of the organizational process in which it is used.

## 4. Shadow AI Governance and Mitigation Strategies

Effective shadow AI governance requires a structured approach that connects organizational policies, employee needs, cybersecurity controls, legal requirements, and ongoing oversight. A complete prohibition on workplace

AI use may appear to provide a simple solution, yet it does not address the productivity, accessibility, and task-related motivations that encourage employees to adopt unauthorized tools. Weak or ambiguous governance creates a different problem by leaving employees to make individual decisions about data sensitivity, acceptable use, and output reliability. Organizations need governance arrangements that make AI use visible, distinguish different levels of risk, provide secure alternatives, and assign responsibility for decisions throughout the AI lifecycle.

Organizational AI governance refers to the rules, practices, processes, and capabilities through which an organization directs and controls its use of AI in accordance with its strategies, values, ethical principles, and legal obligations (Mäntymäki et al., 2022). This definition is relevant to shadow AI because unauthorized use frequently develops in the spaces between formal rules, technological access, and actual working practices. Governance must therefore cover AI systems formally acquired by the organization, employee-selected applications, embedded AI features, custom assistants, external models, and autonomous agents. The scope must also include approved tools that employees use for unapproved data, purposes, or decisions.

### 4.1. Governance Architecture and Risk-Based Oversight

A clear governance architecture provides the organizational foundation for managing shadow AI. Senior management should establish the organization's objectives, risk tolerance, and general principles for AI use. Operational responsibilities can then be distributed among information technology, cybersecurity, data governance, legal affairs, compliance, procurement, human resources, internal audit, and relevant business units. This distribution requires explicit decision rights because fragmented responsibility can allow risky AI practices to remain unaddressed. Mäntymäki et al. (2022) describe organizational AI governance as a system that aligns AI-related decisions with organizational strategy, values, and legal requirements. Shadow AI governance applies this logic to technologies and use cases that may initially remain outside formal decision structures.

A cross-functional AI governance committee can coordinate these responsibilities and prevent governance from becoming the exclusive concern of the IT department. The committee may define acceptable-use rules, evaluate proposed tools, approve higher-risk use cases, review incidents, and monitor changes in legal and technological conditions. Business-unit participation is essential because operational managers understand why employees adopt particular tools and which formal systems fail to meet their needs. Cybersecurity

and legal specialists can evaluate risks that employees or line managers may overlook. Internal audit can assess whether declared policies are reflected in actual practice.

A formal AI policy should define the organizational boundaries of acceptable use. The policy needs to specify which AI services are approved, which data categories may be processed, which activities require prior authorization, and which uses are prohibited. Rules should distinguish public consumer applications from enterprise services that provide contractual protections, administrative controls, and organizational logging. The policy should also cover personal accounts, browser extensions, application programming interfaces, custom assistants, retrieval systems, AI-generated code, and autonomous agents. A policy limited to well-known chatbot websites may become ineffective as AI capabilities are increasingly embedded in ordinary workplace software.

Purpose and context should guide authorization decisions. The same AI service may create a limited risk when used to generate generic brainstorming ideas and a substantial risk when used to process customer records or recommend personnel decisions. The European Union's AI Act similarly adopts a risk-based regulatory approach in which obligations depend on the characteristics, context, and potential effects of an AI system (European Parliament & Council of the European Union, 2024). An organizational framework can adapt this principle by classifying use cases according to data sensitivity, decision impact, system integration, user permissions, affected stakeholders, and degree of autonomy.

A practical classification may separate AI use into low-risk, controlled, high-risk, and prohibited categories. Low-risk uses may include generating generic ideas or editing nonsensitive text. Controlled uses may include working with internal information inside an approved enterprise environment under specified conditions. High-risk uses may involve personal data, confidential information, automated recommendations, external communication, software deployment, or decisions affecting individuals. Prohibited uses may include entering restricted data into public systems, allowing unsupervised AI agents to execute sensitive actions, or using AI for decisions that violate legal or organizational requirements. The purpose of such classification is to match the intensity of governance with the potential severity of harm.

An organizational AI inventory is necessary for applying this classification consistently. The inventory should record approved applications, providers, models, business owners, intended purposes, data categories, integrations, user groups, contractual conditions, and review dates. The NIST AI Risk

Management Framework places governance and contextual mapping at the center of risk management because organizations need to understand where AI is used, who is affected, and which risks arise from each context (National Institute of Standards and Technology [NIST], 2023). An inventory designed for shadow AI should also accept reports of previously unknown tools and uses. Its function is to improve visibility rather than to document only systems that have already completed formal procurement.

A simple and responsive approval process can increase the accuracy of the inventory. Employees are less likely to disclose an AI tool when approval requires lengthy paperwork, unclear communication, or several weeks of waiting. A tiered process can allow rapid approval for low-risk uses while directing sensitive or integrated applications to a more detailed assessment. The assessment should examine the provider, terms of service, privacy conditions, information retention, model-training practices, security controls, technical dependencies, accessibility of logs, and procedures for deleting organizational data. ISO/IEC 42001:2023 supports a management-system approach in which organizations establish, implement, maintain, and continually improve policies and processes for the responsible development, provision, and use of AI systems (International Organization for Standardization [ISO] & International Electrotechnical Commission [IEC], 2023).

AI impact assessments provide a useful mechanism for high-risk use cases. An assessment can document the intended objective, expected benefits, relevant stakeholders, potential harms, data sources, human oversight, limitations, and proposed controls. The assessment should also evaluate whether the use is necessary and whether a less risky method can achieve the same objective. High-impact applications require clear criteria for testing, approval, monitoring, suspension, and retirement. Shadow AI practices that have already become embedded in a business process may require a retrospective assessment before they can be converted into formally governed systems.

Legal and regulatory mapping forms part of the assessment. The organization must determine whether the use involves personal data, protected intellectual property, employment decisions, consumer interactions, financial records, health information, or sector-specific obligations. The AI Act creates responsibilities for providers and deployers according to the role they perform and the risk characteristics of the AI system (European Parliament & Council of the European Union, 2024). Existing data-protection, cybersecurity, contractual, and professional obligations continue to apply when AI is introduced into a process. Organizational approval does not remove these obligations; it provides a mechanism through which they can be identified and addressed.

Risk ownership must remain identifiable after a tool has been approved. Each AI use case should have a business owner responsible for its purpose and outcomes, a technical owner responsible for configuration and integration, and a data owner responsible for the information processed. Higher-risk systems may require independent review from cybersecurity, privacy, legal, or internal-audit functions. Approval decisions should record any accepted residual risks, usage restrictions, review dates, and conditions that would trigger reassessment. Model updates, new integrations, broader user access, and increased autonomy can materially change the original risk profile.

### 4.2. Technical and Operational Controls

Technical controls translate governance decisions into enforceable conditions. Their purpose is to reduce unauthorized data movement, limit excessive permissions, detect unapproved services, preserve records, and prevent AI-generated outputs from producing uncontrolled actions. The NIST Cybersecurity Framework 2.0 organizes cybersecurity outcomes around the functions of Govern, Identify, Protect, Detect, Respond, and Recover (NIST, 2024b). These functions can be adapted to shadow AI by connecting AI governance with existing cybersecurity and incident-management processes.

Approved enterprise AI services provide the first layer of protection. Organizations should offer tools that satisfy common employee needs while providing stronger contractual, administrative, and technical safeguards than personal consumer accounts. Relevant capabilities may include single sign-on, centralized account management, configurable retention, restrictions on model training, organizational logging, access control, and support for data-residency requirements. An approved tool will reduce shadow AI only when it is accessible, functional, and sufficiently responsive to employees' tasks. Haag and Eckhardt (2024) emphasize that effective responses to shadow IT must address cybersecurity requirements and user needs within the same management approach.

Identity and access management should apply the principle of least privilege to AI systems and their integrations. Employees should access only the models, data sources, plugins, and external functions required for their roles. Custom assistants connected to organizational repositories should preserve the access permissions of the underlying documents. Shared accounts should be avoided because they obscure responsibility and make it difficult to revoke access selectively. Privileged AI functions, such as connecting external tools or creating autonomous agents, may require additional approval and stronger authentication.

Data-classification rules should be translated into clear AI-processing rules. Employees need practical guidance showing which data may be entered into public, enterprise, or internally hosted AI systems. Labels such as public, internal, confidential, personal, restricted, and trade secret can be linked to specific processing conditions. Data-loss prevention systems, secure web gateways, endpoint controls, and cloud-access monitoring can help identify or restrict the transfer of sensitive information to unauthorized services. Technical restrictions should be proportionate because overly broad blocking may disrupt legitimate work and encourage employees to seek less visible alternatives.

Prompt and input controls can reduce accidental disclosure within approved systems. Automated checks may detect personal data, credentials, confidential terms, source code, or restricted document labels before information is submitted to a model. Redaction and pseudonymization can reduce exposure when the full identity of a person or organization is unnecessary for the task. These measures do not eliminate the need for employee judgment because automated detection may miss contextual sensitivity. Data minimization should remain the default principle: the model should receive only the information needed to complete the approved task.

Vendor assessment should examine the full AI supply chain rather than focusing exclusively on the visible application. AI services may depend on external model providers, hosting environments, plugins, datasets, libraries, and subcontractors. The assessment should review data retention, secondary use, breach notification, deletion, encryption, access management, change notification, audit rights, service continuity, and dependency management. ISO/IEC 42001:2023 requires organizations to manage AI-related risks and opportunities through a systematic and continually improving management structure (ISO & IEC, 2023). Vendor oversight fits within this structure because external providers can alter the organization's risk exposure throughout the service relationship.

Logging and monitoring should provide evidence about how approved AI systems are used. Relevant records may include user identities, timestamps, selected models, connected data sources, administrative changes, tool calls, generated outputs, and human approvals. The sensitivity of prompts and outputs requires careful decisions about log access and retention. Excessive logging can create a new repository of confidential information, while insufficient logging can prevent incident investigation and accountability. Monitoring should therefore collect information that supports security and audit requirements under clearly defined access controls.

Discovery mechanisms can identify potential shadow AI use without assuming that every detected interaction represents misconduct. Network records, endpoint inventories, browser-extension lists, software-as-a-service discovery tools, procurement data, and expense records may reveal unapproved services. Surveying employees and inviting voluntary disclosure can identify practices that technical monitoring cannot observe. Discovery findings should be evaluated in context because visiting an AI website differs from uploading restricted information or connecting an external agent to an internal database.

Output governance is required because secure input handling does not guarantee reliable results. High-impact outputs should undergo qualified human review before they are used in decisions, external communications, software deployment, or official records. Reviewers need access to relevant source materials and should be able to reject the output without pressure to defer to the model. NIST's Generative Artificial Intelligence Profile identifies confabulation, harmful bias, data privacy, information security, intellectual property, and human overreliance as interconnected areas of generative AI risk (NIST, 2024a). Review procedures should be designed according to the type of harm that could arise from an error.

Provenance records can support review and accountability. Documents, analyses, code, or recommendations produced with substantial AI assistance may need records showing the tool used, date, relevant data sources, reviewer, and level of human modification. The required detail should correspond to the significance of the output. Informal brainstorming does not require the same documentation as a regulatory filing, employment recommendation, or deployed software component. Provenance requirements should preserve traceability without creating administrative burdens that drive AI use back into secrecy.

Agentic AI requires stricter operational boundaries because the system may perform actions across connected environments. Agent permissions should be limited to predefined tools, data sources, and actions. Sensitive operations can require explicit human confirmation, while financial transactions, record deletion, external publication, and privilege changes may remain outside the agent's authority. Sandboxed execution, tool allowlists, rate limits, transaction limits, session timeouts, and emergency termination mechanisms can reduce the consequences of model error or manipulation. OWASP Foundation (2024) identifies excessive agency as a major risk when an AI-enabled application receives functionality, permissions, or autonomy beyond what its task requires.

Security testing should address AI-specific vulnerabilities before deployment. Testing may examine prompt injection, sensitive information disclosure,

insecure output handling, supply-chain weaknesses, model or data poisoning, and excessive agency (OWASP Foundation, 2024). Custom assistants should be tested with adversarial documents and manipulated inputs when they retrieve information from external sources. Generated code should pass ordinary secure-development reviews and automated testing before implementation. AI security must remain connected to established application security because model-related safeguards cannot compensate for weak authentication, insecure interfaces, or excessive user privileges.

Incident-response plans should explicitly include unauthorized AI use. Employees need a clear process for reporting accidental data disclosure, suspicious model behavior, manipulated outputs, compromised plugins, or agent actions. Response procedures should identify who can suspend access, preserve evidence, contact providers, assess affected data, notify stakeholders, and determine legal reporting obligations. NIST's Generative Artificial Intelligence Profile recommends integrating generative AI use cases into incident-response and recovery planning (NIST, 2024a). Existing response teams may require additional expertise to interpret prompts, model configurations, retrieval sources, and AI-generated actions.

Table 3 presents an integrated governance cycle that connects organizational decisions with operational evidence. The cycle begins with discovering actual AI use and continues through classification, authorization, control, monitoring, and organizational learning.

**Table 3** *Integrated Governance Cycle for Managing Shadow AI*

| **Governance function** | **Core practices** | **Expected evidence** | **Principal risks addressed** |
|---|---|---|---|
| Discover | Identify AI tools, embedded features, personal accounts, browser extensions, integrations, custom assistants, and agents used for organizational work | AI inventory, employee disclosures, software discovery records, network and procurement findings | Invisible data flows, undocumented dependencies, unknown vendors |
| Classify | Evaluate data sensitivity, decision impact, affected stakeholders, integration level, user permissions, and system autonomy | Risk classification, data-flow map, preliminary impact assessment | Inconsistent controls, underestimation of high-impact uses, regulatory uncertainty |

| | | | |
|---|---|---|---|
| Authorize | Approve, restrict, reject, or redesign the use case; assign business, technical, and data ownership | Approval record, usage conditions, named owners, review date, accepted residual risks | Fragmented accountability, unauthorized processing, unclear decision rights |
| Enable and protect | Provide approved tools, restrict permissions, apply data controls, test integrations, and establish human-review requirements | Access configuration, vendor assessment, data-processing rules, test results, review procedures | Data leakage, excessive agency, insecure integrations, unreliable outputs |
| Monitor and assure | Review logs, usage patterns, model changes, incidents, output quality, compliance, and control effectiveness | Monitoring reports, audit records, performance indicators, reassessment decisions | Undetected misuse, model drift, policy avoidance, declining control effectiveness |
| Respond and learn | Contain incidents, support affected users, revise controls, improve approved tools, and incorporate useful employee innovations | Incident records, corrective actions, updated policies, formalized use cases, lessons-learned reports | Repeated incidents, persistent shadow use, loss of employee-generated innovation |

***Note. Developed by the author through a synthesis of Haag and Eckhardt (2024), NIST (2023, 2024a, 2024b), ISO and IEC (2023), OWASP Foundation (2024), and Waters-Lynch et al. (2025).***

### 4.3. Employee Enablement and Adaptive Governance

Employee behavior determines whether formal controls will produce secure AI use or drive it further from organizational visibility. Shadow AI often reflects a difference between the technologies employees need and the services the organization provides. Governance should therefore treat employees as participants in risk management and sources of information about emerging use cases. A punitive approach may increase concealment when employees believe that disclosure will result in automatic prohibition or disciplinary action.

Approved alternatives should address the tasks that motivate unauthorized adoption. Employees who need summarization, translation, coding support, document analysis, or brainstorming should have access to suitable enterprise tools and clear guidance on their permitted use. Slow or functionally limited

alternatives may satisfy formal compliance requirements while failing to change actual behavior. Haag and Eckhardt (2024) argue that the management of shadow technology should combine cybersecurity protection with attention to user needs. This balance is central to shadow AI because public generative AI tools often provide immediate functionality with little effort.

AI literacy programs should explain the practical consequences of using different tools and data types. General warnings that AI may be risky are unlikely to guide employees during specific tasks. Training should use realistic scenarios involving customer information, source code, internal reports, meeting transcripts, personnel records, and AI-generated recommendations. Employees need to understand data retention, model limitations, hallucination, prompt injection, intellectual property, personal accounts, and human-review responsibilities. The EU AI Act recognizes AI literacy as part of responsible organizational AI use, placing attention on the knowledge and competence of people who operate AI systems (European Parliament & Council of the European Union, 2024).

Role-specific training is more useful than a uniform course for the entire organization. Software developers need guidance on generated code, external repositories, licensing, and secure testing. Human resources personnel need guidance on personal data, fairness, and decisions affecting employees or applicants. Legal and research staff need guidance on confidentiality, citations, intellectual property, and fabricated sources. Managers need guidance on accountability, appropriate delegation, and the risks of pressuring employees to use AI without formal support.

An accessible disclosure mechanism can surface useful AI practices before they become deeply embedded dependencies. Employees should be able to report a tool, propose a use case, or describe an existing workflow without navigating a complex procurement process. A limited safe-harbor approach may encourage voluntary disclosure when the employee acted to improve a legitimate task and reports the practice before an incident occurs. Deliberate misuse, concealment of known harm, and repeated violation of explicit restrictions can remain subject to ordinary disciplinary procedures. The distinction between good-faith experimentation and reckless behavior strengthens the credibility of governance.

Employee experimentation can be moved into controlled environments. Sandboxes may permit teams to test models with synthetic, anonymized, or low-sensitivity data while the organization evaluates value and risk. Successful experiments can then proceed through formal assessment, integration, and ownership. Waters-Lynch et al. (2025) describe covert generative AI use as

a form of shadow user innovation that may create organizational value while avoiding formal oversight. Governance can capture this value by creating a path through which employee-developed practices become visible, evaluated, and scalable.

Leadership behavior shapes the practical meaning of AI policy. Managers who demand faster AI-enabled results while ignoring approval requirements create conflicting expectations. Senior executives who use personal AI accounts or bypass controls weaken the legitimacy of restrictions applied to other employees. Governance therefore requires consistent conduct across hierarchical levels. Managers should discuss acceptable AI use during project planning and workload allocation rather than treating compliance as a separate technical concern.

Performance measures should evaluate governance quality rather than focusing only on the number of blocked tools or policy violations. Useful measures may include the proportion of AI use cases registered, average approval time, employee access to approved alternatives, frequency of voluntary disclosures, number of high-risk uses remediated, completion of role-specific training, AI-related incidents, and recurrence of previously identified problems. A declining number of detected tools may indicate improved compliance, reduced detection capacity, or deeper concealment. Measures should therefore be interpreted alongside employee feedback and technical evidence.

Continuous review is necessary because AI services, organizational uses, legal obligations, and threat techniques change rapidly. An approved tool may introduce new features, connect to additional data sources, change its retention terms, or add autonomous functions. A low-risk use case may become high-risk when it expands to new users or begins influencing consequential decisions. ISO/IEC 42001:2023 places continual improvement within the structure of an AI management system (ISO & IEC, 2023). Shadow AI governance should adopt the same orientation through scheduled reassessment, incident learning, and policy revision.

Adaptive governance creates a controlled path from discovery to legitimate organizational use. Low-value and high-risk shadow practices can be discontinued. Useful practices can be redesigned, secured, and formally supported. Uncertain practices can remain within limited experimentation environments until adequate evidence is available. This approach treats shadow AI as a source of security exposure and organizational learning. Its effectiveness depends on maintaining visibility, proportionality, accountability, and a credible response to employee needs.

## 5. Future Research Directions

Shadow AI has recently emerged as a distinct research subject at the intersection of management information systems, cybersecurity, organizational behavior, knowledge management, and AI governance. Existing studies have clarified its basic characteristics, identified several cybersecurity risks, and emphasized the tension between employee-driven innovation and organizational control (Puthal et al., 2025; Waters-Lynch et al., 2025). Research has also begun to examine governance responses and the relationship between shadow AI use and organizational knowledge leakage (Chin et al., 2025; Dolci & Aguiar, 2025). The available evidence remains limited in comparison with the speed and diversity of organizational AI adoption. Many current arguments are based on conceptual analysis, practitioner observations, or insights transferred from the established shadow IT literature. A stronger evidence base requires clear conceptual boundaries, validated measurement instruments, longitudinal research designs, multilevel analysis, and direct evaluations of governance interventions.

Future research should preserve the dual character of shadow AI. A narrow focus on security violations may overlook productivity gains, employee learning, local experimentation, and the discovery of unmet technological needs. A purely innovation-oriented perspective may underestimate data leakage, unreliable outputs, accountability gaps, and the expansion of organizational attack surfaces. Chin et al. (2025) found an inverted U-shaped relationship between shadow AI use and organizational knowledge leakage within metaverse-related businesses, indicating that its consequences may change according to the intensity and context of use. Waters-Lynch et al. (2025) similarly frame covert generative AI use as a form of employee-driven innovation whose organizational value depends on whether firms can identify and integrate useful practices. These findings support a research agenda that investigates conditions, thresholds, and trade-offs rather than assuming uniformly beneficial or harmful outcomes.

### 5.1. Conceptualization, Measurement, and Multilevel Explanation

Conceptual clarity represents the first research priority. Studies need to distinguish shadow AI from approved enterprise AI, employee-initiated AI, business-managed AI, accidental policy violations, malicious AI misuse, and the unauthorized use of approved systems. A technology may be selected by an employee and remain aligned with organizational governance after disclosure and approval. An approved application may enter the shadow domain when employees use restricted data, activate unapproved integrations, or apply its

outputs to unauthorized decisions. Future definitions should therefore address the status of the tool, the visibility of the practice, the data involved, the purpose of use, the degree of system integration, and the level of autonomy granted to the AI.

The unit of analysis also requires clarification. Shadow AI can refer to an individual act, a recurring employee behavior, a team-level routine, an unofficial technological system, or an organizational condition characterized by limited AI visibility. These levels should not be combined without explanation. An employee who occasionally uses a public chatbot for nonsensitive editing presents a different theoretical and practical case from a department that connects an external model to customer records through an undocumented application programming interface. Research designs should identify whether they examine shadow AI adoption, usage frequency, concealment, risk intensity, organizational prevalence, or dependence on unauthorized AI-supported processes.

Validated measurement instruments are needed to support cumulative empirical research. A binary measure asking whether an employee has used an unapproved AI tool cannot represent the range of relevant behaviors. A multidimensional scale could assess undisclosed tool adoption, unauthorized data processing, policy-inconsistent use of approved systems, hidden AI integration, reliance on unverified outputs, and delegation of actions to unauthorized agents. Frequency, duration, organizational importance, data sensitivity, and system autonomy could be measured separately because they influence the severity of a practice. Scale development should begin with qualitative interviews, critical-incident reports, expert review, and employee diaries before proceeding to exploratory and confirmatory validation across independent samples.

Measurement research must address social desirability and concealment. Employees may underreport shadow AI use when they fear sanctions, reputational damage, or managerial disapproval. Managers may also underestimate its prevalence because they interpret the absence of reported incidents as evidence of compliance. Anonymous surveys can reduce this problem but may remain vulnerable to self-presentation and recall errors. Randomized-response techniques, list experiments, indirect questioning, and scenario-based measures may produce more accurate estimates of sensitive behavior. Digital trace data from network records, browser extensions, software inventories, and AI service logs could be combined with employee surveys when legal, ethical, and privacy conditions permit.

Behavioral studies should examine why employees choose shadow AI when approved alternatives exist. Performance expectancy, effort expectancy, task–technology misfit, workload, time pressure, AI self-efficacy, perceived policy legitimacy, and dissatisfaction with official systems may shape this decision. Social influences may arise when colleagues and supervisors normalize AI use without discussing authorization or data protection. Employees may also rely on neutralization strategies that allow them to interpret policy violations as harmless, necessary, or beneficial to the organization. Statements such as "everyone uses it," "the data were not truly confidential," or "the organization did not provide a suitable tool" may reduce perceived personal responsibility.

Several theoretical perspectives can explain different parts of this behavior. Task–technology fit can explain how deficiencies in approved systems encourage employees to seek external tools. Technology acceptance perspectives can clarify the influence of perceived usefulness and ease of use. Deterrence theory can examine the effects of sanction certainty and severity, while protection motivation theory can address threat perceptions and coping responses. Neutralization theory can explain how employees justify policy-inconsistent behavior. Affordance theory can show how accessibility, concealability, generativity, and automation create opportunities for unauthorized use. A sociotechnical systems perspective can connect individual behavior with work design, organizational structures, security controls, and technological capabilities.

The coexistence of these perspectives does not require placing many theories within a single model. Research should select the theoretical lens that matches the focal question and level of analysis. A study examining initial adoption may focus on task–technology fit and performance expectancy. A study examining concealment may use neutralization, perceived policy legitimacy, or psychological safety. A governance study may examine deterrence, organizational justice, and trust. A multilevel study may connect leadership signals and security climate with individual decisions. Theoretical precision will provide greater explanatory value than models containing numerous weakly connected constructs.

Outcome research should extend beyond data leakage. Potential employee-level outcomes include productivity, learning, creativity, job autonomy, stress, role ambiguity, and perceived surveillance. Team-level outcomes may include knowledge sharing, workflow adaptation, coordination, and the development of undocumented dependencies. Organizational outcomes may include information security incidents, innovation speed, decision quality, compliance costs, intellectual property exposure, resilience, and AI capability

development. Chin et al. (2025) demonstrate that shadow AI may have nonlinear relationships with knowledge-related outcomes. Future studies should test curvilinear, threshold, and configurational effects across a broader range of organizational settings.

Temporal dynamics deserve greater attention. Shadow AI practices may begin as occasional experiments and gradually become established routines. The perceived usefulness of a tool may increase employee reliance, while repeated use may lead to the transfer of more sensitive information or broader integration with internal systems. Organizational responses may also change behavior over time. An initial prohibition may reduce visible use but increase concealment, whereas access to a suitable enterprise alternative may gradually redirect employees toward approved services. Longitudinal surveys, diary studies, repeated digital-trace observations, and process studies can capture these transitions more effectively than cross-sectional designs.

Multilevel research can explain how organizational conditions shape individual behavior. Policy clarity, ethical climate, information security culture, AI literacy programs, leadership conduct, availability of approved tools, and the responsiveness of IT departments may influence employees' risk assessments. Team-level norms may create variation within the same organization because some managers actively discuss acceptable AI use while others focus exclusively on performance outcomes. Cross-level models could examine whether organizational governance reduces shadow AI directly or changes the effects of task pressure, AI self-efficacy, and perceived usefulness on employee behavior.

Sectoral and institutional contexts should be incorporated into theory testing. Healthcare, finance, education, public administration, software development, consulting, and research organizations differ in data sensitivity, professional responsibility, regulatory exposure, and acceptable tolerance for experimentation. A shadow AI practice that creates limited consequences in generic content preparation may be unacceptable in clinical, legal, financial, or employment decisions. Comparative studies can identify whether the same drivers and governance responses operate consistently across sectors.

Cross-national research can examine how legal systems, cultural values, labor relations, and institutional trust affect shadow AI. Employees' willingness to disclose unauthorized use may vary according to power distance, perceptions of managerial fairness, job security, and confidence in organizational reporting procedures. Data-protection regimes and AI regulations may also affect policy design and managerial attention. Studies based on a single country should

avoid treating their findings as universal, particularly when organizational AI use is shaped by different regulatory and cultural environments.

### 5.2. Governance Effectiveness, Technical Detection, and Agentic AI

Governance research should move from descriptive recommendations toward comparative evaluation. Existing work proposes policies, inventories, employee training, approved alternatives, technical monitoring, and controlled experimentation environments as potential responses to shadow AI (Dolci & Aguiar, 2025; Puthal et al., 2025). The effectiveness of these measures remains an empirical question. Organizations may adopt similar policies while producing different outcomes because of implementation quality, employee trust, managerial consistency, and the usefulness of approved systems.

Future studies should compare restrictive, enabling, and adaptive governance models. A restrictive model emphasizes blocking, sanctions, and centralized approval. An enabling model focuses on approved tools, guidance, and rapid access to AI capabilities. An adaptive model combines risk-based controls with sandboxes, voluntary disclosure, continuous review, and pathways for formalizing useful employee innovations. Comparative case studies can reveal how these models operate in practice, while longitudinal designs can assess their effects on security incidents, employee behavior, innovation, and organizational trust.

Field experiments and natural experiments would strengthen causal evidence. Organizations may introduce training, approved AI platforms, revised policies, safe-harbor reporting procedures, or new monitoring technologies in phases. Researchers could compare business units before and after implementation or examine matched units receiving different interventions. Relevant outcomes may include the use of personal AI accounts, attempts to transfer sensitive data, voluntary disclosure rates, time required to obtain approval, employee satisfaction with authorized tools, and the number of useful shadow practices converted into governed applications.

Governance effectiveness should not be measured solely through reductions in detected unauthorized use. A decrease may indicate stronger compliance, reduced AI use, displacement to personal devices, or more successful concealment. Evaluation should combine security indicators with employee perceptions, technical records, incident data, and operational outcomes. Suitable indicators may include policy comprehension, perceived legitimacy, trust in the reporting process, use of approved alternatives, approval speed,

recurrence of violations, decision quality, productivity, and the rate at which employee innovations become formally supported.

The relationship between monitoring and employee privacy requires systematic investigation. Technical discovery mechanisms can identify visits to AI services, browser extensions, unapproved integrations, and potential transfers of sensitive data. Extensive monitoring may create concerns about workplace surveillance, autonomy, and the collection of employee prompts. Employees who perceive monitoring as disproportionate may avoid organizational systems or reduce voluntary disclosure. Research should examine which forms of monitoring employees consider legitimate, how transparency affects acceptance, and whether privacy-preserving discovery methods can provide sufficient security visibility.

Technical studies should evaluate the accuracy and limitations of shadow AI detection. Web traffic classification may identify known services while missing embedded AI features, locally hosted models, encrypted connections, personal devices, or newly created applications. Data-loss prevention systems may detect recognizable personal data or confidential labels but fail to understand contextual sensitivity. False positives can interrupt legitimate work and weaken trust in security systems. False negatives may create misplaced confidence. Benchmark datasets and standardized evaluation criteria are needed to compare detection approaches under realistic organizational conditions.

Privacy-preserving monitoring presents a promising research area. Organizations need information about risky AI use without creating repositories containing every employee prompt and output. Future systems could apply local classification, data minimization, pseudonymization, aggregate reporting, differential privacy, or risk-based escalation. Low-risk events might be recorded only as aggregated usage patterns, while high-risk transfers could trigger more detailed review under defined procedural safeguards. Studies should evaluate the security value, employee acceptability, and legal implications of these architectures.

Research should also investigate the conversion of shadow AI into governed organizational capability. Some employee-created workflows may offer substantial value once their data sources, permissions, ownership, and review requirements have been formalized. Waters-Lynch et al. (2025) argue that covert generative AI use may contribute to capability renewal when organizations can identify and integrate valuable employee innovations. Process research could examine how organizations discover these practices, decide which ones to support, transfer individual knowledge to formal teams, and prevent dependence on undocumented personal systems.

Agentic AI creates an urgent extension of the shadow AI research agenda. Employees can increasingly configure systems that retrieve information, call external tools, communicate with other services, and execute multistep tasks. Unauthorized agent use changes the risk profile because an error or manipulated instruction can produce an organizational action. Research should distinguish shadow AI tools that generate advice from shadow agents that exercise delegated authority.

Agentic AI studies should examine permission design, action limits, human confirmation, identity management, tool access, memory, and inter-agent communication. Low-code agent development may allow employees to create complex automations without fully understanding the permissions or dependencies involved. Emerging research on agentic explainability reports limited organizational visibility into agent configurations and interactions across agent networks (Elsayed & Jones, 2026). Shadow agent research should therefore consider observability at design time and runtime, including who created the agent, which resources it can access, what decisions it makes, and how its actions can be reconstructed.

Human oversight requires more precise operationalization in agentic environments. Requiring a person to click "approve" may provide weak protection when the user lacks time, expertise, or sufficient information to evaluate the proposed action. Future experiments should compare forms of oversight, including confirmation before every action, approval for specific risk categories, transaction limits, exception-based review, dual authorization, and retrospective auditing. Research should measure how these designs affect error detection, employee workload, automation benefits, and responsibility attribution.

Liability and accountability research will become increasingly important as shadow agents affect customers, employees, and external stakeholders. An employee may configure the agent, a third-party provider may operate the model, an organizational account may supply access, and a manager may benefit from its output. The distribution of control across these actors complicates responsibility. Legal, information systems, and organizational behavior researchers should jointly examine how organizations assign ownership and how employees understand their responsibility when AI systems perform actions on their behalf. Table 4 summarizes a research agenda that connects these gaps with suitable questions, levels of analysis, and methodological approaches.

***Table 4 Research Agenda for Shadow AI***

| **Research domain** | **Illustrative research questions** | **Primary level of analysis** | **Suitable methods** |
|---|---|---|---|
| Conceptual boundaries | When does employee-initiated AI become shadow AI? How should unauthorized tools be distinguished from unauthorized uses of approved systems? | Practice, system, organization | Concept analysis, expert panels, Delphi studies, comparative case studies |
| Measurement | Which dimensions capture the frequency, concealment, sensitivity, integration, and autonomy of shadow AI use? | Individual and team | Interviews, scale development, validation studies, randomized-response techniques |
| Behavioral drivers | How do task–technology misfit, work pressure, AI self-efficacy, social norms, and policy legitimacy influence shadow AI? | Individual | Surveys, experiments, diary studies, multilevel modeling |
| Innovation and performance | Under which conditions does shadow AI improve productivity, learning, creativity, or capability development? | Individual, team, organization | Longitudinal surveys, field studies, process research, configurational analysis |
| Security and knowledge outcomes | How do usage intensity, data sensitivity, integration, and autonomy shape data leakage and decision risk? | Use case and organization | Incident analysis, digital traces, nonlinear modeling, simulations |
| Governance effectiveness | Which combinations of policies, approved alternatives, training, disclosure mechanisms, and sanctions produce sustainable compliance? | Team and organization | Field experiments, natural experiments, comparative cases, difference-in-differences analysis |
| Employee trust and surveillance | How does AI-use monitoring affect privacy concerns, trust, disclosure, and concealment? | Individual and organization | Vignette experiments, surveys, interviews, privacy impact assessments |
| Technical discovery | How accurately can organizations detect embedded, local, encrypted, and agent-based shadow AI while limiting false alarms? | System and network | Benchmarking, security testing, prototype evaluation, red-team exercises |
| Formalization of employee innovation | How can valuable shadow practices be transferred into secure and scalable organizational systems? | Team and organization | Longitudinal case studies, action research, process tracing |

| | | | |
|---|---|---|---|
| Shadow agents | How do autonomy, memory, permissions, and inter-agent communication alter security and accountability? | Agent, workflow, organization | Agent simulations, controlled experiments, runtime log analysis, threat modeling |
| Sectoral and cross-national variation | How do regulation, professional norms, culture, and data sensitivity change the causes and consequences of shadow AI? | Sector and country | Cross-country surveys, comparative institutional analysis, multigroup models |

The research domains in Table 4 are interconnected. Measurement quality will influence the validity of studies examining behavioral drivers and organizational outcomes. Governance research requires reliable indicators of actual behavior, while technical detection research must consider employee trust and privacy. Agentic AI research will also require cooperation among cybersecurity, information systems, organizational behavior, law, and human–computer interaction scholars. Interdisciplinary designs should retain a clear focal question and define how each disciplinary perspective contributes to its explanation.

Future evidence should ultimately help organizations answer three practical questions: where shadow AI exists, why employees rely on it, and which response produces an acceptable balance among security, accountability, innovation, and employee needs. Research that addresses these questions through transparent measurement, longitudinal evidence, and field-based evaluation can move the literature from general risk awareness toward a mature body of organizational knowledge.

## 6. Conclusion

Shadow AI has emerged as a significant organizational challenge as employees gain direct access to powerful generative and agentic AI tools. These technologies can support productivity, creativity, learning, and rapid problem-solving, yet their use outside formal organizational oversight creates serious risks for information security, privacy, intellectual property, regulatory compliance, and decision quality. The defining problem is the gap between the AI systems organizations believe they manage and the tools, data flows, integrations, and automated processes that employees actually use.

The analysis presented in this chapter shows that shadow AI cannot be explained solely as employee misconduct or weak cybersecurity. Its emergence is shaped by technological accessibility, performance expectations, task–technology mismatches, time pressure, inadequate organizational tools,

ambiguous policies, and social norms surrounding AI use. Employees often adopt unauthorized AI applications because these tools provide immediate solutions to operational problems. Shadow AI therefore reflects weaknesses in organizational technology provision and governance as well as individual choices.

The risks associated with shadow AI extend across the confidentiality, integrity, and availability of organizational information. Employees may expose confidential data, personal information, source code, internal documents, and intellectual property to unapproved external services. AI-generated outputs may introduce factual errors, fabricated information, bias, insecure code, or misleading recommendations into organizational workflows. Undocumented integrations and autonomous agents may expand the organizational attack surface by connecting external models to internal data sources and business applications. These risks become more difficult to detect and manage when organizations lack visibility into how AI is actually being used (National Institute of Standards and Technology [NIST], 2024; Puthal et al., 2025).

Effective governance requires a balance between organizational control and employee enablement. Blanket prohibitions may reduce visible use while encouraging employees to move their AI activities to personal accounts, devices, and less observable environments. Weak governance leaves employees to make complex decisions about data sensitivity, legal requirements, and output reliability without sufficient guidance. A more effective approach combines clear policies, risk-based classification, approved enterprise tools, rapid authorization procedures, technical safeguards, human review, employee training, monitoring, and incident-response mechanisms. Haag and Eckhardt (2024) similarly argue that shadow technology should be managed through an approach that addresses cybersecurity requirements and legitimate user needs.

Organizational AI governance should treat visibility as a central objective. Organizations need to identify which AI tools are used, what data they process, how they are integrated into workflows, which decisions they influence, and who is responsible for their outcomes. AI inventories, impact assessments, ownership structures, access controls, data-classification rules, vendor evaluations, and provenance records can support this objective. Governance arrangements should remain adaptive because AI applications, provider conditions, threat techniques, and organizational use cases change rapidly. The governance of AI is therefore an ongoing organizational capability rather than a one-time compliance exercise (Mäntymäki et al., 2022; NIST, 2023).

Employee participation is essential for improving organizational visibility. Workers are often the first to recognize where AI can improve a process or

where existing systems fail to meet operational requirements. Reporting mechanisms, controlled experimentation environments, and pathways for formalizing useful employee-created solutions can help organizations convert valuable shadow practices into secure organizational capabilities. Waters-Lynch et al. (2025) emphasize that covert generative AI use may contribute to organizational renewal when employee innovations are identified, evaluated, and integrated into formal structures. Governance should therefore distinguish responsible experimentation from reckless or harmful use.

The growth of agentic AI will make this distinction increasingly important. AI systems are beginning to move beyond content generation toward task execution, tool use, data retrieval, and automated interaction with business systems. Unauthorized agents may act through employee accounts, access multiple data sources, and perform actions without adequate human review. Future governance models must therefore consider levels of autonomy, permission boundaries, identity management, tool access, action limits, and accountability for AI-generated decisions and actions.

Shadow AI ultimately represents a test of organizational adaptability. Organizations that respond through rigid control may suppress useful experimentation without eliminating hidden use. Organizations that ignore the issue may expose themselves to escalating security, legal, and operational risks. Sustainable governance requires proportionate controls, secure alternatives, credible accountability, and continuous dialogue between employees, managers, technical specialists, legal professionals, and security teams.

The central argument of this chapter is that shadow AI should be managed through visibility, proportionality, accountability, and enablement. Visibility allows organizations to understand actual AI use. Proportionality ensures that controls correspond to the sensitivity and impact of each use case. Accountability clarifies responsibility for data, systems, outputs, and decisions. Enablement provides employees with secure and practical alternatives to unauthorized tools. Organizations that develop these capabilities will be better positioned to benefit from AI while protecting their information assets and maintaining trust among employees, customers, regulators, and other stakeholders.

## References

Barberá, I. (2025). *AI privacy risks & mitigations: Large language models (LLMs)*. European Data Protection Board.

Barlette, Y., Berthevas, J.-F., Richet, J.-L., & Georg Schaffner, L. (2025). Investigating the influence of emotions on shadow IT usage behaviours. *Systèmes d'Information & Management, 30*(2), 99–149. https://doi.org/10.54695/sim.252.0099

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium* (pp. 2633–2650). USENIX Association. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

Chin, T., Li, Q., Mirone, F., & Papa, A. (2025). Conflicting impacts of shadow AI usage on knowledge leakage in metaverse-based business models: A Yin-Yang paradox framing. *Technology in Society, 81*, Article 102793. https://doi.org/10.1016/j.techsoc.2024.102793

Dolci, P. C., & Aguiar, M. S. (2025). Governance and generative artificial intelligence: Challenges and risks of shadow AI in business environment. In *Proceedings of the 31st Americas Conference on Information Systems (AMCIS 2025)*. Association for Information Systems. https://aisel.aisnet.org/amcis2025/lacais/lacais/1/

Elsayed, Y., & Jones, C. (2026). Agentic explainability at scale: Between corporate fears and XAI needs [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2604.14984

European Parliament, & Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Union, L 119*, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

European Parliament, & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union, L*, 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*, 625–630. https://doi.org/10.1038/s41586-024-07421-0

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering, 66*, 111–126. https://doi.org/10.1007/s12599-023-00834-7

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security* (pp. 79–90). Association for Computing Machinery. https://doi.org/10.1145/3605764.3623985

Haag, S., & Eckhardt, A. (2017). Shadow IT. *Business & Information Systems Engineering, 59*(6), 469–473. https://doi.org/10.1007/s12599-017-0497-x

Haag, S., & Eckhardt, A. (2024). Dealing effectively with shadow IT by managing both cybersecurity and user needs. *MIS Quarterly Executive, 23*(4), 399–412. https://doi.org/10.17705/2msqe.00104

International Organization for Standardization, & International Electrotechnical Commission. (2023). *Information technology—Artificial intelligence—Management system* (ISO/IEC Standard No. 42001:2023). International Organization for Standardization. https://www.iso.org/standard/81230.html

Klotz, S., Kopper, A., Westner, M., & Strahringer, S. (2019). Causing factors, outcomes, and governance of shadow IT and business-managed IT: A systematic literature review. *International Journal of Information Systems and Project Management, 7*(1), 15–43. https://doi.org/10.12821/ijispm070102

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics, 2*(4), 603–609. https://doi.org/10.1007/s43681-022-00143-x

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

National Institute of Standards and Technology. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1

National Institute of Standards and Technology. (2024b). *The NIST cybersecurity framework (CSF) 2.0* (NIST CSWP 29). U.S. Department of Commerce. https://doi.org/10.6028/NIST.CSWP.29

Nguyen, T. (2024). Understanding shadow IT usage intention: A view of the dual-factor model. *Online Information Review, 48*(3), 500–522. https://doi.org/10.1108/OIR-04-2022-0243

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

OWASP Foundation. (2024). *OWASP Top 10 for LLM applications 2025*. https://genai.owasp.org/llm-top-10/

Puthal, D., Mishra, A. K., Mohanty, S. P., Longo, A., & Yeun, C. Y. (2025). Shadow AI: Cyber security implications, opportunities and challenges in the unseen frontier. *SN Computer Science, 6*, Article 405. https://doi.org/10.1007/s42979-025-03962-x

Waters-Lynch, J., Allen, D. W. E., Potts, J., & Berg, C. (2025). Shadow user innovation: Governing covert generative-AI use for dynamic-capability renewal. *Innovation: Organization & Management*, 1–17. https://doi.org/10.1080/14479338.2025.2519546