Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 12 Issue 5, May - 2025

Shadow AI in the Workplace: A Socio-Technical Governance and Risk Management Framework

Ozuomba Simeon

TETFUND Center of Excellence in Computational Intelligence Research,
University of Uyo
Akwa Ibom State , Nigeria
simeonozuomba@uniuyo.edu.ng, simeonoz@yahoo.com

Abstract

Shadow AI refers to the use of artificial intelligence tools for organizational work without formal authorization, visibility or control. Its growth reflects a widening gap between employees' demand for fast AI-enabled productivity and organizations' capacity to approve, monitor and govern AI use. This paper develops a socio-technical governance and risk management framework for Shadow AI through an integrative conceptual review of literature on Shadow IT, technology workarounds, technology acceptance, responsible AI governance, data governance, cybersecurity and human-AI interaction. The analysis shows that Shadow AI differs from traditional Shadow IT because it hides not only tools and data flows, but also inference, generation, classification, recommendation and delegated action. The paper identifies adoption drivers, develops a risk taxonomy, proposes a risk-prioritization and control-maturity model, and presents a seven-layer governance framework covering accountability, policy and use-case rules, inventory and discovery, data classification, technical security controls, human oversight and audit-response mechanisms. The study contributes to AI governance research by translating responsible AI principles into operational controls for everyday workplace use. It concludes that prohibition alone is too weak as a governance posture; organizations should reduce the friction of approved AI use while making high-risk AI activity visible, assessable and accountable.

Keywords: Shadow AI; generative AI; AI governance; Shadow IT; data governance; cybersecurity; risk management; human oversight; workplace automation

1. Introduction

Recent workplace evidence confirms that AI use is spreading faster than many formal governance processes. The 2024 Microsoft and LinkedIn Work Trend Index reported that 75% of knowledge workers used generative AI at work and that 78% of AI users brought their own AI tools to work [1]. McKinsey's 2024 global survey also reported a sharp increase in organizational generative-AI use, with 65% of respondents saying their organizations regularly used generative AI [2]. These figures do not prove that every workplace AI use is unsafe, but they show why organizations need practical controls for AI activity that may begin at the employee level before enterprise approval catches up.

Generative artificial intelligence has moved from specialist computing environments into ordinary workplace routines. Employees now use AI tools to draft emails, summarize meetings, prepare reports, write code, translate text, build spreadsheets, review contracts and explore data. These tools can raise productivity, but they can also produce plausible errors, expose sensitive prompts, generate biased outputs and create new cybersecurity entry points. The risks are therefore socio-technical: they emerge from the interaction of people, data, vendors, work pressure, model behaviour and organizational controls [3,4].

The governance challenge is intensified by the speed and convenience of public AI services. A worker can open a chatbot, install an AI browser extension or connect an AI meeting assistant long before a formal procurement review, privacy assessment or security test is completed. Risk frameworks such as the NIST AI Risk Management Framework and the NIST Generative AI Profile encourage organizations to govern, map, measure and manage AI

risk across context-specific use cases, but those functions are difficult to apply when the AI tool is unknown to the organization [5,6].

This gap has produced Shadow AI: the use of AI systems for work without formal authorization, visibility or control. The phenomenon is related to Shadow IT, where employees adopt unapproved digital tools to overcome delays, rigid systems or poorly supported work needs [7]. It also reflects workaround behaviour, because employees often depart from prescribed processes when approved systems do not help them complete a legitimate task efficiently [8]. Shadow AI, however, is more complex than conventional Shadow IT because it hides model-mediated inference, generation, recommendation and action, not only unapproved storage or communication.

Shadow AI matters because its harms may be difficult to reconstruct. The organization may not know what data was submitted, whether the vendor retained it, what output was produced, how the output influenced a decision or whether a human checked it. A confidential report pasted into a public chatbot, a coding assistant receiving proprietary source code, or an AI plug-in with broad access to email and files can create confidentiality, privacy, intellectual property, regulatory and cybersecurity problems. In decision-support contexts, hidden AI use can also affect fairness, accountability and trust.

This paper addresses three research questions: RQ1: What organizational and technological factors drive Shadow AI adoption in the workplace? RQ2: What governance, cybersecurity, privacy, compliance and accountability risks arise from Shadow AI? RQ3: What control framework can organizations use to detect, govern and safely enable workplace AI use?

The paper makes three contributions. First, it conceptualizes Shadow AI as a distinct form of unauthorized workplace technology use that combines the hidden adoption logic of Shadow IT with the distinctive risks of AI inference, generation and automation. Second, it develops a risk taxonomy that links Shadow AI practices to organizational harms and control responses. Third, it proposes a seven-layer socio-technical governance framework that translates responsible AI principles into practical controls for ordinary workplace use.

2. Literature Review

2.1 Shadow IT, workarounds and unauthorized technology use

Shadow AI is best understood against the background of Shadow IT. Shadow IT research shows that unauthorized systems are not always caused by carelessness; they frequently emerge where formal systems do not meet work needs, where procurement is slow, or where users find official tools too rigid for urgent tasks [7]. Hidden technology use is therefore both a risk and a signal. It reveals where organizational controls and practical work conditions have fallen out of alignment.

Workaround theory deepens this view. Workarounds occur when users intentionally depart from prescribed processes to achieve a goal that formal systems do not support well [8]. In AI-enabled workplaces, the workaround may be a chatbot that rewrites a customer complaint, a coding assistant that receives source code, a browser plug-in connected to corporate mail, or an AI agent acting across documents. Shadow AI inherits the organizational logic of workarounds but increases the scale, speed and opacity of the resulting risk.

2.2 Adoption drivers: usefulness, ease of use and transformation pressure

Technology acceptance research helps explain why employees adopt AI tools even where permission is unclear. Perceived usefulness and perceived ease of use strongly shape information-system adoption [9]. Later models emphasize performance expectancy, effort expectancy, social influence and facilitating conditions [10]. Generative AI scores highly on these dimensions because it offers immediate help with writing, coding, summarization and analysis.

Digital transformation research also shows that organizations face pressure to redesign processes and capabilities around digital technologies [11]. Where formal transformation lags behind employee expectations, workers may improvise with consumer tools already available through browsers or personal accounts. In such settings, vague prohibitions carry little practical force unless organizations provide safe alternatives that are easy enough to use.

2.3 Responsible AI governance and accountability

Responsible AI governance seeks to convert principles such as fairness, transparency, privacy, accountability, safety and human control into organizational arrangements. Recent governance reviews emphasize that responsible AI requires structural, relational and procedural mechanisms that specify who is accountable, what is governed and how controls are maintained [12]. The global ethics literature shows substantial convergence around high-level principles, but implementation remains uneven when principles are not connected to roles, processes and evidence [13,14].

Operational work on AI ethics has therefore moved from statements of intent to tools, methods and governance practices [15]. This shift is important for Shadow AI because a general AI policy is not enough when employees use tools outside approved channels. Governance must be able to identify tools, classify use cases, control data movement, document approvals and support review of outputs.

Accountability is especially difficult in AI because agency is distributed across users, vendors, models, prompts, data pipelines and business processes. Algorithmic auditing can narrow this accountability gap when it creates evidence of what was done, why it was allowed and how risk was controlled [16,17]. In Shadow AI settings, however, the evidence trail is often absent. Discovery, logging, approved environments and human validation become core governance requirements rather than administrative additions.

2.4 Data governance, privacy and cybersecurity

Shadow AI is also a data governance problem. Data governance defines decision rights and accountabilities for data-related processes [18]. Modern data governance further emphasizes roles, standards, monitoring and value creation [19]. When employees paste internal data into unapproved AI systems, these decision rights are bypassed: the data owner may not know that the data has left controlled systems, and the privacy team may not have assessed lawful basis, minimization, retention or cross-border transfer risk.

Cybersecurity research on language models shows that AI systems can be manipulated, misused and integrated insecurely. Indirect prompt injection can compromise LLM-integrated applications by hiding malicious instructions in content later processed by the model [20]. Prompt-injection work also shows that language models may follow hostile instructions embedded in prompts or context [21]. More broadly, the malicious-use literature warns that AI can lower the cost of deception, social engineering, vulnerability discovery and targeted manipulation [22]. These risks are amplified when unvetted AI tools receive corporate data or obtain access to enterprise resources.

2.5 Fairness, explainability and human-AI interaction

The risks of Shadow AI are not limited to confidentiality. AI-generated content can embed bias, hide assumptions and influence decisions in ways that are difficult to reconstruct. Fairness research warns that technical abstractions may miss the institutional and social context in which automated systems operate [23]. The debate on explanation and automated decision-making also shows why accountability becomes difficult when algorithmic outputs are opaque or poorly documented [24].

Explainable AI research has responded by calling for systems and processes that make AI outputs interpretable to human decision-makers [25]. Human-AI interaction guidelines emphasize transparency, control, error recovery, uncertainty communication and appropriate reliance [26]. These concerns are directly relevant to Shadow AI because employees may rely on outputs from tools that the organization has not assessed, documented or approved.

2.6 Research gap

The literature provides strong building blocks, but they remain fragmented. Shadow IT explains hidden adoption; workaround theory explains user improvisation; technology acceptance explains motivation; responsible AI explains principles; data governance explains accountabilities; cybersecurity research explains technical threats; and human-AI interaction explains reliance and oversight. What remains underdeveloped is an integrated workplace framework that focuses on informal, employee-driven and partly invisible AI use. Existing AI governance models often assume that the AI system is known and can be assessed through formal lifecycle controls. Shadow AI begins from the opposite condition: the organization may not know that the tool, prompt,

plug-in, data connection or AI-assisted decision exists. This paper fills that gap by developing a framework that links hidden employee adoption to enterprise controls.

Taken together, these studies explain important parts of the Shadow AI problem, but they do not fully address the workplace condition in which AI tools are useful, accessible, partly invisible and capable of influencing decisions without formal approval. Shadow IT literature explains hidden tool adoption, but not hidden inference, generation and delegated action. Responsible-AI literature explains principles and lifecycle governance, but often assumes that the AI system is known to the organization. Cybersecurity research explains technical attack surfaces, but is less explicit about how everyday workarounds create those exposures. This gap justifies a Shadow-AI-specific framework that links employee work practices to governance accountability, data control, technical security, human oversight and audit evidence.

3. Methodological Approach

The study uses an integrative conceptual review and framework-development approach. This approach is appropriate where the research problem cuts across several mature and emerging domains and where the goal is to build a practical model rather than estimate a single empirical relationship. The review was organized around six knowledge areas: Shadow IT and workarounds, technology acceptance, digital transformation, responsible AI governance, data governance and AI cybersecurity.

The source selection process followed a transparent protocol. Searches were organized around Scopus, Web of Science, ACM Digital Library, IEEE Xplore, ScienceDirect, SpringerLink, Emerald, Taylor & Francis, Google Scholar, NIST publications and ISO/IEC standards. The search covered peer-reviewed and standards-oriented sources published before 2025, with emphasis on literature relevant to information systems, computing, management, cybersecurity, data governance, responsible AI and workplace technology use.

The search terms combined conceptual and technical phrases, including Shadow IT, technology workarounds, generative AI governance, responsible AI governance, organizational AI governance, AI risk management, algorithmic accountability, data governance, prompt injection, human-AI interaction, workplace AI adoption and AI audit. The review used backward and forward citation checking where a source was central to the argument. The final synthesis retained 39 sources, including peer-reviewed articles, conference papers, books, formal standards and authoritative industry or government reports that directly informed the framework.

Sources were included when they addressed unauthorized technology use, workplace adoption, AI governance, accountability, privacy, cybersecurity, human oversight or organizational control. Sources were excluded when they focused only on model performance, algorithmic optimization or technical architecture without relevance to workplace governance. Standards and guidance documents were included where they provided operational language for AI management, risk assessment or generative-AI-specific risk controls, especially NIST and ISO/IEC materials [5,6,27,28].

The synthesis was thematic. Each source was examined for the problem it addressed, the governance gap it implied and the control response it supported. Themes were grouped into adoption drivers, risk categories, policy mechanisms, technical controls, data-governance requirements, oversight practices and audit-response needs. The final corpus retained 39 sources because each item made a direct contribution to one or more of these synthesis domains.

Table 1. Source selection and synthesis protocol

Stage	Purpose	Operational decision	Output
Domain identification	Capture knowledge relevant to Shadow AI	Use search blocks covering Shadow IT, workarounds, technology acceptance, responsible AI, data governance and AI cybersecurity.	Initial domain corpus
Traceability screening	Improve source reliability	Prioritize peer-reviewed sources, recognized standards and publications with stable publisher metadata or durable identifiers.	Screened source set
Conceptual relevance	Avoid unrelated AI literature	Retain sources explaining unauthorized use, governance, risk, accountability, data handling, security or oversight.	Governance-relevant corpus
Thematic coding	Derive framework domains	Code sources for adoption drivers, risk categories, visibility gaps, data controls, technical controls, human review and audit needs.	Thematic map
Framework synthesis	Produce an actionable model	Consolidate recurring control responses into mutually reinforcing layers aligned with workplace implementation.	Seven-layer framework
Review flow	Document screening and synthesis	Retain sources that directly address workplace adoption, AI governance, data control, cybersecurity, accountability, standards or human oversight; use citation chasing for central works.	Final synthesis corpus of 39 sources.

Source: Author's conceptual synthesis.

4. Conceptualizing Shadow AI

4.1 Definition and scope

Shadow AI is defined in this paper as the use of AI tools, models, plug-ins, agents or AI-enabled services for organizational work without formal authorization, visibility or control. The definition covers public chatbots, AI-powered browser extensions, meeting transcription tools, code generation assistants, AI analytics services, image generators, automated workflow agents and AI features embedded in third-party software.

A tool becomes Shadow AI not because it is inherently unsafe, but because the organization lacks approval, inventory, data-use rules, vendor review, technical monitoring, human oversight, logging or accountability. A sanctioned enterprise AI assistant can also become Shadow AI if a department connects it to sensitive repositories without approval or uses it for a prohibited decision. Conversely, personal experimentation with public information may be low risk when it does not involve organizational data, systems, customers or decisions.

Shadow AI is not necessarily malicious. It should be distinguished from malicious insider use because many cases begin as unauthorized but work-oriented attempts to complete legitimate tasks. Malicious use involves deliberate harm, theft, deception or sabotage. The distinction matters because governance should encourage safe disclosure and better support while still enforcing clear boundaries for protected information and decisions with material consequences.

4.2 Shadow AI versus Shadow IT

Shadow AI overlaps with Shadow IT but should not be treated as identical. The difference lies in the kind of organizational action being hidden. Shadow IT often hides infrastructure, storage, communication or workflow tools. Shadow AI may hide inference, generation, classification, summarization, prediction and delegated action. Table 2 summarizes the distinction.

Table 2. Distinguishing Shadow IT and Shadow AI

Dimension	Shadow IT	Shadow AI
Typical function	Unauthorized storage, communication, workflow or analytics tool.	Unauthorized inference, generation, recommendation, coding, summarization or automation.
Primary hidden asset	Software, cloud service, data store or device.	Model, prompt, output, agent, plug-in, data connection or AI-assisted decision.
Main risk emphasis	Security, integration, compliance and data management.	Security, data leakage, hallucination, bias, accountability, IP exposure and decision influence.
Governance problem	IT visibility and control.	AI visibility, data control, model behaviour, human validation and accountability.
Audit challenge	What system was used and what data was stored.	What data was submitted, what output was generated, who relied on it and how it affected work.

Source: Author's conceptual synthesis.

4.3 Categories and escalation of Shadow AI use

Shadow AI can be grouped into five workplace categories: content assistance, knowledge extraction, software and technical work, decision support and delegated action. The risk level usually rises when AI use moves from public information to confidential data, from drafting to decision influence, and from passive assistance to autonomous action. This progression matters because many employees begin with low-risk drafting and gradually move toward higher-risk uses without a formal review point.

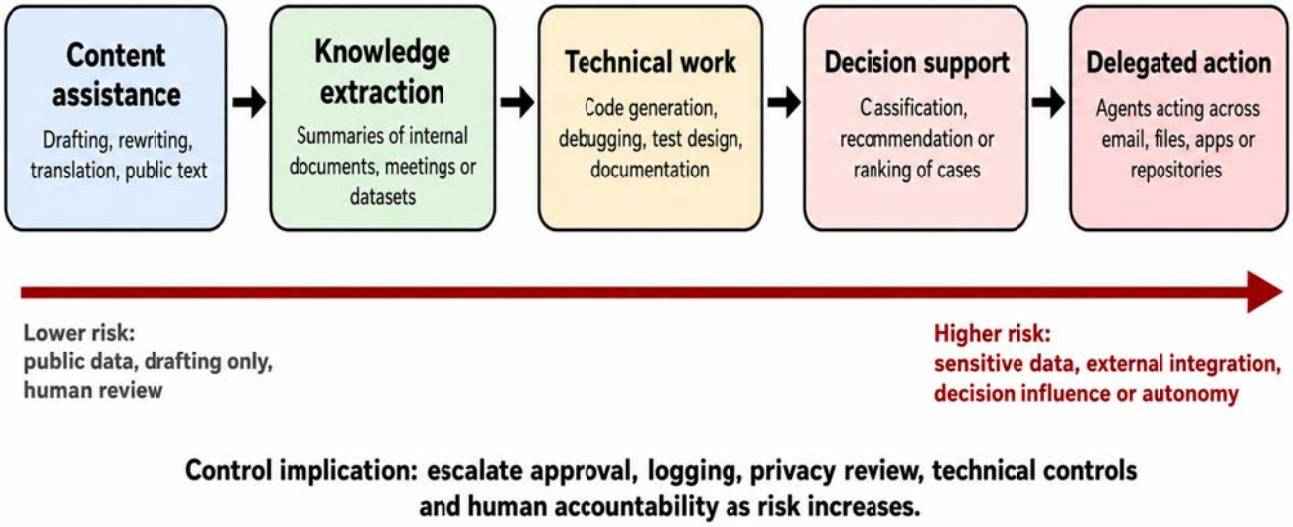

Figure 1. Risk escalation path from low-risk AI assistance to high-risk autonomous Shadow AI activity.

5. Drivers of Shadow AI in the Workplace

The drivers of Shadow AI are mainly organizational and operational. Employees face time pressure, communication overload, large document volumes and rising expectations for speed. AI tools offer immediate relief. Where approved tools are absent, difficult to access or poorly matched to work needs, employees often reach for consumer AI systems. Table 3 links the main drivers to theory and governance response.

Table 3. Drivers of Shadow AI and governance implications

Driver	Theoretical basis	Workplace example	Governance implication
Productivity pressure	Workaround theory	Employee uses a public chatbot to compress a long report before a deadline.	Provide approved AI assistants for routine drafting and summarization.
Perceived usefulness	Technology acceptance	Staff experience immediate gains in writing, coding or analysis.	Design approved tools around real work tasks, not only compliance language.
Ease of access	Technology acceptance	Consumer AI tools are available through browser, phone or personal account.	Use allowlists, browser controls and clear routes to approved alternatives.
Weak approved alternatives	Digital transformation gap	Official tools are slow, unavailable or hard to request.	Shorten approval cycles and provide departmental AI sandboxes.
Unclear policy	Governance gap	Employee is unsure whether internal notes can be summarized by AI.	Use task-based policy categories: permitted, restricted and prohibited.
Fear of disclosure	Organizational culture	Employee hides AI use because it may be judged as laziness or misconduct.	Create a safe reporting culture while enforcing clear boundaries for sensitive data.

Source: Author's conceptual synthesis.

These drivers show why blanket prohibition rarely works. If an organization only blocks tools without addressing work pressure and unmet needs, employees may move AI use to personal devices, private accounts or unsupervised plug-ins. Governance must therefore combine restrictions on high-risk behaviour with approved routes for legitimate productivity needs.

6. Governance Challenges and Organizational Risks

6.1 Governance challenges

The foundational challenge is visibility. If AI use is hidden, the organization cannot classify the tool, assess the data, validate the output or investigate incidents. Visibility failure then creates accountability failure: the organization may not know who used the tool, what information was submitted, what was generated or how the output influenced work.

A second challenge is policy translation. Many AI policies state broad principles but do not tell employees what to do when they need to summarize internal meeting notes, draft a customer response, debug code or analyze a dataset. A third challenge is ownership. Shadow AI crosses IT, cybersecurity, legal, privacy, procurement, internal audit, human resources and business operations. Without clear ownership, departments may assume another unit is responsible.

A fourth challenge is culture. If employees believe every AI use will be treated as misconduct, they may conceal experimentation. If the organization treats every AI use as innovation, unsafe practice may become normalized. Mature governance creates a climate where employees can ask for guidance, while enforcing clear boundaries for sensitive data, regulated activity and high-impact decisions.

6.2 Risk taxonomy

The main risks of Shadow AI can be grouped into eleven categories: data leakage, privacy violation, intellectual property exposure, unreliable output, bias and discrimination, cybersecurity exposure, regulatory non-compliance, loss of auditability, provenance failure, vendor/data-residency risk and organizational knowledge contamination. Table 4 links each risk to a typical workplace scenario and a control response.

Table 4. Shadow AI risk-control matrix

Risk	Typical Shadow AI scenario	Likely harm	Primary control response
Data leakage	Confidential report pasted into a public chatbot.	Loss of confidentiality and contractual breach.	Data classification, DLP and approved enterprise AI environment.
Privacy violation	Customer or staff records uploaded for summarization.	Unlawful processing and privacy exposure.	Privacy review, lawful-basis check, minimization and access control.
IP exposure	Developer submits proprietary code or patentable design.	Loss of trade secret or ownership ambiguity.	Coding-assistant governance, secret scanning and vendor review.
Unreliable output	AI-generated analysis copied into a management report.	False decisions, rework and reputational damage.	Human review, evidence checks and output validation.

Risk	Typical Shadow AI scenario	Likely harm	Primary control response
Bias and discrimination	AI used informally for recruitment screening.	Unfair treatment and legal exposure.	High-risk-use prohibition unless formally assessed.
Cybersecurity exposure	Browser extension requests broad access to corporate mail.	Credential abuse, data exfiltration and account compromise.	Extension control, OAuth governance and least privilege.
Regulatory non-compliance	AI used in regulated decisions without documentation.	Audit failure and legal penalty.	Risk classification, approval workflow and documentation.
Loss of auditability	No record of prompts, outputs or review steps.	Weak accountability and poor incident investigation.	Logging, case records and AI-use registers.
Provenance failure	AI fabricates citations, legal authorities or technical evidence.	False reporting, poor decisions and loss of credibility.	Source verification, citation checks and output labelling.
Vendor and residency risk	Unapproved AI service stores or processes data in unknown locations.	Contractual, privacy and cross-border transfer risk.	Vendor due diligence, data-processing terms and residency controls.
Knowledge contamination	AI-generated errors, fabricated references or false policy interpretations enter official documents or repositories.	Institutional error, poor decisions and reputational damage.	Evidence checks, source verification, approval workflow and periodic content review.

Source: Author's conceptual synthesis.

6.3 Cybersecurity threat scenarios

Cybersecurity risk increases when AI tools are connected to enterprise resources. A public chatbot used only for public text drafting may be low risk. A browser extension with access to mail, calendars, documents and identity tokens is different: it can become a privileged channel for data exposure or account compromise. Table 5 summarizes common threat scenarios.

Table 5. Cybersecurity threats associated with Shadow AI

Threat	Shadow AI scenario	Control
Prompt injection	Employee uses AI to summarize external content containing hidden malicious instructions.	Approved tools, content isolation, output checks and user warnings.
Sensitive prompt disclosure	Confidential data entered into a public AI system.	DLP, prompt monitoring, classification rules and training.
OAuth abuse	AI plug-in requests access to email or cloud files.	App consent governance, least privilege and administrator approval.
API key leakage	Developer shares source code containing secrets with an AI coding tool.	Secret scanning, repository controls and approved coding assistant.
Over-permissioned agents	AI agent acts across files and applications without review.	Agent sandboxing, approval gates, logging and rollback procedure.
RAG poisoning	Untrusted documents are indexed into an AI retrieval system and influence outputs.	Source vetting, retrieval controls, content signing and monitoring.
Supply-chain risk	Unvetted AI extension or SaaS AI feature processes work files.	Vendor due diligence, extension allowlist and contract review.

Source: Author's conceptual synthesis.

6.4 Regulatory and standards implications

Shadow AI creates compliance uncertainty because it may bypass data protection impact assessments, vendor due diligence, contractual review and sector-specific controls. The highest concern arises when AI affects people in employment, finance, education, healthcare, public services, legal, security or safety-sensitive processes. In such settings, informal AI use can create undocumented automated decision support even where the user regards the AI output as only a helpful suggestion.

A standards-oriented approach is therefore useful. The NIST AI RMF provides a broad structure for governing, mapping, measuring and managing AI risk [5], while the NIST Generative AI Profile identifies generative-AI-specific concerns such as confabulation, information integrity, data exposure, harmful bias and misuse [6]. ISO/IEC 42001 provides requirements and guidance for an AI management system [27], and ISO/IEC 23894 provides guidance for AI-related risk management [28]. These materials support a risk-based approach in which low-risk drafting is handled differently from sensitive data processing, high-impact decisions and autonomous action.

The framework should be adapted to jurisdictional requirements. Data protection, employment, consumer protection, intellectual-property and sector-specific rules vary across countries. A control approach that is acceptable for low-risk document drafting may be insufficient where AI use affects employment, finance, education, healthcare, legal services, public administration or regulated personal data.

7. Risk Prioritization and Control Maturity

A practical Shadow AI program needs a simple way to prioritize use cases before deeper legal, privacy or security review. The scoring approach below supports early triage. It should not replace expert assessment; it helps managers decide which use cases require escalation, logging, formal approval or prohibition.

For each Shadow AI use case i, the risk priority score can be calculated as a weighted score:

$$RPS_i = w_S S_i + w_L L_i + w_E E_i + w_C C_i + w_A A_i \quad (1)$$

$$s.t. w_S + w_L + w_E + w_C + w_A = 1 \quad (2)$$

where RPS_i is the risk priority score for use case i, S_i is data sensitivity, L_i is likelihood of exposure, E_i is external integration risk, C_i is compliance impact and A_i is the autonomy level of the AI-assisted action. Each factor may be scored from 1 to 5. The weighting coefficients are non-negative and sum to 1, allowing organizations to reflect sector-specific risk appetite. A hospital, for example, may weight privacy and compliance more heavily; a software company may weight intellectual property and code exposure more heavily.

Control maturity can also be expressed as a simple index:

$$CMI = \frac{G + P + I + D + T + H + R}{7} \quad (3)$$

where CMI is the control maturity index, G is governance accountability, P is policy maturity, I is inventory and discovery maturity, D is data classification maturity, T is technical-control maturity, H is human oversight maturity and R is audit-response maturity. Each component can be scored from 1 to 5. Equation (3) helps management identify whether the organization is relying too much on policy while neglecting technical monitoring, training or audit. A CMI value closer to 1 indicates weak or immature Shadow AI control capability, while a value closer to 5 indicates a more mature and balanced governance environment. The index should not be interpreted as a precise statistical measure; it is a managerial screening tool for comparing maturity across control domains.

Table 6. Shadow AI risk classification matrix

Risk level	Indicative score	Example use	Required response
Low	1.0-1.9	Drafting public text from public information.	Permit with guidance and disclosure rules.
Moderate	2.0-2.9	Summarizing internal but non-sensitive notes.	Use approved AI tool and keep human review.
High	3.0-3.9	Processing confidential business documents.	Require data-owner approval, DLP and logging.
Very high	4.0-4.5	AI used in HR, finance, legal or customer-impacting decisions.	Formal risk review, legal/privacy assessment and senior approval.
Critical	4.6-5.0	Public AI tool receives secrets, credentials, source code or regulated personal data.	Prohibit, contain and assess as a security or privacy incident.

Source: Author's conceptual synthesis.

8. Theoretical Framing: Shadow AI as a Socio-Technical Control Failure

Shadow AI is a socio-technical control failure because it arises from the interaction of people, tools, tasks, incentives and governance structures. The technical affordances of AI tools matter: they are accessible, conversational and powerful. Organizational context matters just as much: employees face deadlines, approved systems may be slow, and policies may not reflect daily work. Risk is produced by the mismatch among these elements.

Three theoretical lenses explain the problem. Shadow IT and workaround theory explain why employees bypass official channels when formal systems do not support real work. Technology acceptance explains why generative AI is attractive: it is useful, easy to use and often available without institutional friction. Socio-technical control theory explains why risk cannot be managed by technical blocking alone; incentives, norms, data rules, accountability and human judgment must be aligned.

Foundation models and large language models intensify this alignment problem because reusable general-purpose capabilities can be adapted across many work tasks [29,30]. Transformer architectures make these systems flexible in processing contextual inputs and generating fluent outputs [31]. Management research further shows that AI is both a strategic tool and an organizational design challenge, because automation and augmentation often coexist in practice [32,33]. In high-impact domains such as human resources, informal AI use requires particular caution [34].

The control model therefore treats AI output as a form of data work that must be governed, interpreted and validated. Data-mining literature shows that analytical outputs depend on data quality, modelling assumptions and interpretation choices [35]. Explainable AI and AI user-experience research reinforce the need for outputs that people can interrogate and understand [36,37]. Automation-bias research warns that humans may over-trust machine output in time-pressured settings [38]. Ethical principles remain essential, but principles become useful only when they are translated into organizational routines, evidence and controls [39].

The seven-layer framework follows from this framing. Governance accountability addresses ownership; policy rules address formal norms; inventory and discovery address visibility; data classification addresses information boundaries; technical controls address system-level enforcement; human oversight addresses user judgment; and audit-response mechanisms address learning and accountability.

9. Proposed Governance and Risk Management Framework

The proposed framework contains seven layers. Each layer responds to a recurring problem identified in the literature and risk analysis. The layers are not intended to operate as a linear checklist only; they reinforce one another through feedback. Audit findings should refine policy, training, discovery and technical controls. Employee reports should improve the inventory. Incidents should update risk scoring and approval rules.

Table 7. Derivation of the seven-layer Shadow AI governance framework

Literature-derived problem	Governance implication	Framework layer
Unauthorized technology use and workarounds.	Organizations must understand why employees bypass official channels.	Governance accountability.
Unclear acceptable use.	Employees need practical rules for ordinary work tasks.	Policy and use-case rules.
Lack of visibility.	Unknown tools cannot be assessed or controlled.	AI inventory and discovery.
Uncontrolled movement of sensitive data.	AI use must be linked to data sensitivity and privacy obligations.	Data classification and privacy.
Prompt injection, plug-ins, API keys and access abuse.	AI controls must be integrated into cybersecurity architecture.	Technical security controls.
Overreliance on AI output.	Humans need skills and responsibility to validate outputs.	Human oversight and literacy.
Weak evidence trail and poor incident learning.	AI use must be auditable and improvable.	Audit, monitoring and response.

Source: Author's conceptual synthesis.

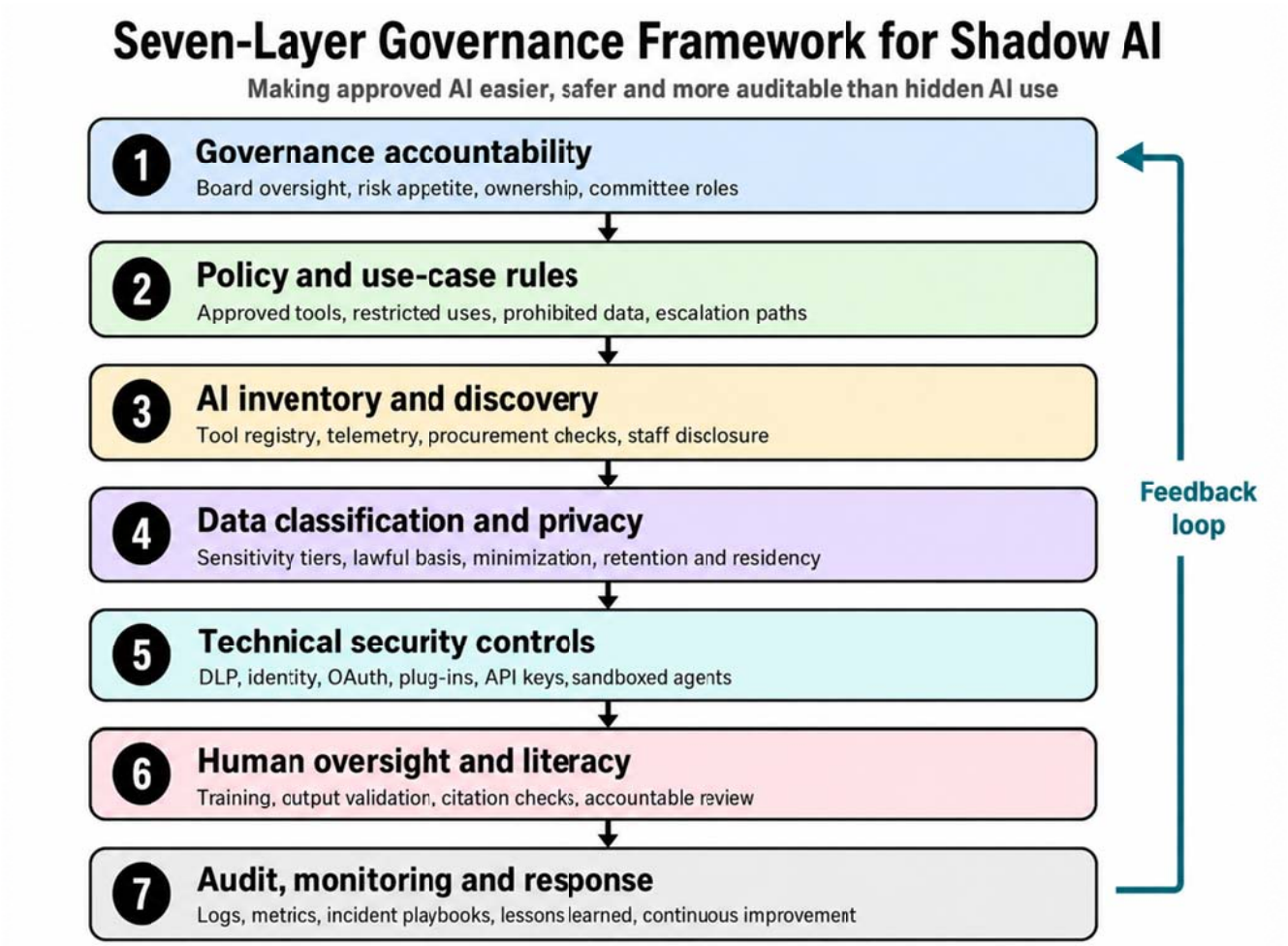

Figure 2. Seven-layer socio-technical governance framework for managing Shadow AI in workplace environments.

9.1 Layer 1: Governance accountability

The organization should assign clear ownership for AI governance. A practical structure includes executive sponsorship, an AI governance committee and defined roles for IT, cybersecurity, legal, privacy, procurement, internal audit, human resources and business units. The objective is not to centralize every small AI decision, but to ensure that risk appetite, approval routes and accountability are clear.

9.2 Layer 2: Policy and use-case rules

Policies should translate principles into daily choices. Employees should know which AI tools are approved, which data must not be entered, when disclosure is required and when a use case must be escalated. A useful policy distinguishes between permitted, restricted and prohibited use rather than treating all AI activity as the same.

9.3 Layer 3: AI inventory and discovery

An AI inventory should include approved AI systems, embedded AI functions, browser extensions, plug-ins, APIs, departmental tools and AI-enabled vendor products. Discovery should combine technical monitoring with self-disclosure, procurement review and periodic departmental surveys. The aim is to reduce invisibility without creating a culture of suspicion.

9.4 Layer 4: Data classification and privacy

AI use should be tied to data classification. Public information may be suitable for low-risk tools. Internal data may require approved enterprise environments. Confidential, regulated or highly sensitive data should require explicit approval, contractual safeguards and documented review. This layer turns data protection from an abstract concern into an operational rule.

9.5 Layer 5: Technical security controls

Technical controls should include data loss prevention, identity-based access, single sign-on, browser extension governance, OAuth consent controls, API key management, endpoint monitoring, logging, plug-in restrictions, approved tool allowlists and sandboxing for experimental AI agents. Controls should be proportionate: the organization should not over-control low-risk drafting, but it should strongly control high-risk data movement and autonomous action.

9.6 Layer 6: Human oversight and AI literacy

Employees need practical AI literacy. Training should use real workplace scenarios: summarizing customer data, using AI to write code, checking AI-generated references, reviewing AI-drafted legal language and recognizing prompt-injection risks. Human oversight is especially important when AI output influences decisions, customers, legal positions, safety, financial reporting or employment outcomes.

9.7 Layer 7: Audit, monitoring and response

Audit should verify whether approved tools are used, whether sensitive data is protected, whether outputs are reviewed and whether incidents are recorded. Incident response should cover accidental prompt disclosure, unauthorized upload of confidential files, malicious plug-ins, compromised AI accounts, hallucinated official content and over-permissioned agents. Lessons from incidents should feed back into policy, training and technical controls.

10. Implementation Roadmap

Organizations need a staged path because Shadow AI cannot be solved in one policy cycle. Table 8 proposes a practical roadmap with owners, evidence and indicative timing. The phases may overlap, but the sequence helps organizations move from visibility to classification, governance, control and continuous improvement.

Table 8. Implementation roadmap for Shadow AI governance

Phase	Main objective	Key actions	Responsible actors	Evidence of completion	Indicative timing
1. Discover	Find where AI is already being used.	Survey employees; review extensions, procurement records and AI-enabled vendors.	IT, CISO, procurement, business units.	Initial AI inventory and risk heat map.	First 30-60 days
2. Classify	Separate low-risk from high-risk use.	Classify data, tools and use cases; flag sensitive departments and regulated processes.	Data owners, legal, privacy, cybersecurity.	Risk classification register.	60-90 days
3. Govern	Create accountable rules.	Approve policy, assign owners, define escalation and publish tool guidance.	Executive sponsor, AI governance committee, HR.	Approved policy and governance charter.	90-120 days
4. Control	Make safe use practical and unsafe use difficult.	Deploy approved tools, DLP, identity controls, plug-in restrictions, logging and vendor review.	IT, CISO, procurement, vendors.	Control checklist and monitoring dashboard.	120-180 days
5. Improve	Learn from use and incidents.	Audit AI use, review incidents, update training and refresh the inventory.	Internal audit, governance committee, business owners.	Quarterly governance review and improvement plan.	Quarterly

11. Framework Alignment and Evaluation Logic

The framework can be evaluated analytically by checking whether it covers the main risk-management functions expected in recognized AI governance materials. Table 9 maps the proposed layers to the NIST AI RMF functions and ISO/IEC management-system logic. The mapping shows that the framework is not a replacement for those standards; it adapts their logic to the specific problem of informal and partly invisible workplace AI use.

The framework can also be tested through scenario-based evaluation. In a public-chatbot scenario, the key controls are data classification, user guidance and output checking. In an AI coding-assistant scenario, the key controls are vendor review, secret scanning, code-retention rules and human review. In an AI plug-in or agent scenario, the key controls are OAuth governance, least privilege, sandboxing, logging and rollback. These scenarios show that the same seven layers apply at different intensity levels depending on data sensitivity, integration depth, compliance impact and autonomy.

This alignment provides analytical validation of coverage, but it does not replace empirical validation through case studies, expert panels, surveys, Delphi analysis or design-science evaluation.

Table 9. Alignment of the proposed framework with recognized AI governance logic

Proposed layer	NIST AI RMF alignment	ISO/IEC management-system alignment	Practical control value
Governance accountability	Govern	Leadership, roles, policy and objectives.	Assigns ownership and escalation authority.
Policy and use-case rules	Govern / Map	Operational planning and documented procedures.	Defines permitted, restricted and prohibited use.
AI inventory and discovery	Map	Asset and process identification.	Makes hidden tools and data connections visible.
Data classification and privacy	Map / Measure	Risk assessment and data controls.	Links AI use to data sensitivity and privacy duties.
Technical security controls	Manage	Operational control and risk	Controls plug-ins, identity,

Proposed layer	NIST AI RMF alignment	ISO/IEC management-system alignment	Practical control value
		treatment.	DLP, APIs and agents.
Human oversight and literacy	Measure / Manage	Competence, awareness and operational control.	Reduces overreliance through training and review.
Audit, monitoring and response	Manage / Govern	Performance evaluation and continual improvement.	Supports monitoring, evidence and improvement.

Source: Author's conceptual synthesis based on NIST and ISO/IEC governance logic.

12. Discussion

12.1 Response to research questions

RQ1 asked what drives Shadow AI adoption. The analysis shows that Shadow AI is driven by productivity pressure, perceived usefulness, ease of access, lack of approved alternatives, unclear policies, low AI literacy and fear of disclosing AI use. These drivers are consistent with technology acceptance and workaround theory: employees adopt tools that help them complete tasks when formal systems do not meet their needs.

RQ2 asked what risks arise from Shadow AI. The paper identifies risks in eleven areas: data leakage, privacy violation, intellectual property exposure, unreliable output, bias and discrimination, cybersecurity exposure, regulatory non-compliance, loss of auditability, provenance failure, vendor/data-residency risk and organizational knowledge contamination. The risk grows when AI use involves sensitive data, external integrations, high-impact decisions or autonomous actions.

RQ3 asked what framework organizations can use. The proposed seven-layer framework provides an integrated answer by connecting governance accountability, policy, discovery, data classification, technical security, human oversight and audit. The framework is designed to make AI use traceable and accountable without suppressing legitimate innovation.

12.2 Theoretical contribution

The paper extends Shadow IT literature by showing that unauthorized AI use changes the nature of hidden technology risk. Shadow IT often concerns unapproved tools and data flows; Shadow AI adds hidden inference, hidden generation, hidden recommendation and hidden decision support. The paper also contributes to responsible AI governance by moving from principles to workplace controls. Fairness, accountability and transparency become practical only when organizations know which tools are used, what data is submitted, how outputs are validated and who is responsible.

12.3 Practical contribution and governance trade-offs

The central practical contribution is governed enablement. Employees use Shadow AI because it solves real work problems. Organizations should therefore provide approved AI pathways, clear rules and fast support. At the same time, they should control sensitive data movement, high-impact decisions, external integrations and autonomous actions. The governance goal is to reduce the appeal of hidden AI by making approved AI safer and easier to use.

The framework also recognizes important trade-offs. Excessive monitoring can damage trust and push experimentation underground. A strict ban can create a false sense of control while employees continue using personal accounts. Uncontrolled flexibility can expose confidential data and weaken accountability. Small organizations may lack resources for all controls at once. For that reason, the proposed model uses proportionality: low-risk drafting should be guided, high-risk data movement should be controlled, and AI-assisted decisions about people or regulated activities should be formally assessed.

13. Managerial Implications

Managers should begin by treating Shadow AI as an enterprise risk rather than a purely technical issue. The first question is not which website to block, but which work problems employees are trying to solve. A mature response combines discovery, education, approved tools, data protection and accountability.

Boards and executives should define risk appetite for AI use and require periodic reporting on high-risk use cases. CIOs and CISOs should maintain tool inventories, access controls, extension governance and monitoring. Legal and privacy teams should review sensitive use cases, vendor terms and cross-border data issues. Human resources should integrate AI literacy into staff training and employment-related controls. Internal audit should test whether controls operate in practice. Department heads should identify where AI is useful and where it is too risky.

A practical rule for employees is simple: if the information would not be sent to an external party without approval, it should not be entered into an unapproved AI tool. This rule is easy to remember, while formal policy can handle more complex cases involving lawful basis, confidential data, regulated work and high-impact decisions.

14. Conclusion

Shadow AI is now a practical reality of workplace digital life. Employees use AI because it helps them work, but hidden AI use can weaken data protection, cybersecurity, accountability, compliance and trust. The challenge is not to eliminate AI experimentation, but to bring it within a governable structure.

This paper has argued that Shadow AI should be understood as a socio-technical control failure and as a signal of unmet organizational needs. The proposed seven-layer framework gives organizations a way to respond without relying on blanket restriction. By combining accountability, policy, discovery, data classification, technical controls, human oversight and audit, organizations can make AI use safer, traceable and accountable.

The strongest governance posture is disciplined enablement: employees should have approved routes for legitimate AI-assisted work, while high-risk activity involving sensitive data, regulated decisions, external integrations or autonomous action should be visible, assessed and controlled. That balance gives organizations a practical path between unmanaged enthusiasm and ineffective prohibition.

References

[1] Microsoft and LinkedIn, 2024 Work Trend Index Annual Report: AI at Work Is Here. Now Comes the Hard Part. Microsoft Corporation, 2024. https://news.microsoft.com/annual-wti-2024/

[2] McKinsey & Company, The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value. McKinsey & Company, 2024. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024

[3] L. Weidinger et al., "Ethical and social risks of harm from language models," arXiv, 2021. https://doi.org/10.48550/arXiv.2112.04359

[4] Y. K. Dwivedi et al., "So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," International Journal of Information Management, vol. 71, 102642, 2023. https://doi.org/10.1016/j.ijinfomgt.2023.102642

[5] E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, 2023. https://doi.org/10.6028/NIST.AI.100-1

[6] A. S. Shah et al., Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. National Institute of Standards and Technology, 2024. https://doi.org/10.6028/NIST.AI.600-1

[7] M. Silic and A. Back, "Shadow IT: A view from behind the curtain," Computers & Security, vol. 45, pp. 274-283, 2014. https://doi.org/10.1016/j.cose.2014.06.007

[8] S. Alter, "Theory of workarounds," Communications of the Association for Information Systems, vol. 34, article 55, 2014. https://doi.org/10.17705/1CAIS.03455

[9] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Quarterly, vol. 13, no. 3, pp. 319-340, 1989. https://doi.org/10.2307/249008

[10] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," MIS Quarterly, vol. 27, no. 3, pp. 425-478, 2003. https://doi.org/10.2307/30036540

[11] G. Vial, "Understanding digital transformation: A review and a research agenda," Journal of Strategic Information Systems, vol. 28, no. 2, pp. 118-144, 2019. https://doi.org/10.1016/j.jsis.2019.01.003

[12] M. Mäntymäki, M. Minkkinen, T. Birkstedt, and M. Viljanen, "Defining organizational AI governance," AI and Ethics, vol. 2, pp. 603-609, 2022. https://doi.org/10.1007/s43681-022-00143-x

[13] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, pp. 389-399, 2019. https://doi.org/10.1038/s42256-019-0088-2

[14] L. Floridi et al., "AI4People - An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," Minds and Machines, vol. 28, pp. 689-707, 2018. https://doi.org/10.1007/s11023-018-9482-5

[15] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices," Science and Engineering Ethics, vol. 26, pp. 2141-2168, 2020. https://doi.org/10.1007/s11948-019-00165-5

[16] I. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency, pp. 33-44, 2020. https://doi.org/10.1145/3351095.3372873

[17] J. Mökander, M. Morley, M. Taddeo, and L. Floridi, "Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations," Science and Engineering Ethics, vol. 29, article 49, 2023. https://doi.org/10.1007/s11948-023-00460-0

[18] V. Khatri and C. V. Brown, "Designing data governance," Communications of the ACM, vol. 53, no. 1, pp. 148-152, 2010. https://doi.org/10.1145/1629175.1629210

[19] R. Abraham, J. Schneider, and J. vom Brocke, "Data governance: A conceptual framework, structured review, and research agenda," International Journal of Information Management, vol. 49, pp. 424-438, 2019. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

[20] K. Greshake et al., "Not what you have 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, 2023. https://doi.org/10.1145/3605764.3623985

[21] F. Perez and I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 3030-3050, 2022. https://doi.org/10.18653/v1/2022.emnlp-main.225

[22] M. Brundage et al., "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," arXiv, 2018. https://doi.org/10.48550/arXiv.1802.07228

[23] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59-68, 2019. https://doi.org/10.1145/3287560.3287598

[24] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," International Data Privacy Law, vol. 7, no. 2, pp. 76-99, 2017. https://doi.org/10.1093/idpl/ipx005

[25] A. Rai, "Explainable AI: From black box to glass box," Academy of Management Discoveries, vol. 6, no. 1, pp. 137-141, 2020. https://doi.org/10.5465/amd.2020.0046

[26] S. Amershi et al., "Guidelines for human-AI interaction," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, article 3, 2019. https://doi.org/10.1145/3290605.3300233

[27] ISO/IEC, ISO/IEC 42001:2023 Artificial intelligence - Management system. International Organization for Standardization and International Electrotechnical Commission, 2023. https://www.iso.org/standard/42001

[28] ISO/IEC, ISO/IEC 23894:2023 Artificial intelligence - Guidance on risk management. International Organization for Standardization and International Electrotechnical Commission, 2023. https://www.iso.org/standard/77304.html

[29] R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv, 2021. https://doi.org/10.48550/arXiv.2108.07258

[30] T. B. Brown et al., "Language models are few-shot learners," arXiv, 2020. https://doi.org/10.48550/arXiv.2005.14165

[31] A. Vaswani et al., "Attention is all you need," arXiv, 2017. https://doi.org/10.48550/arXiv.1706.03762

[32] A. Kaplan and M. Haenlein, "Siri, Siri, in my hand: Who is the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence," Business Horizons, vol. 62, no. 1, pp. 15-25, 2019. https://doi.org/10.1016/j.bushor.2018.08.004

[33] S. Raisch and S. Krakowski, "Artificial intelligence and management: The automation-augmentation paradox," Academy of Management Review, vol. 46, no. 1, pp. 192-210, 2021. https://doi.org/10.5465/amr.2018.0072

[34] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," California Management Review, vol. 61, no. 4, pp. 15-42, 2019. https://doi.org/10.1177/0008125619867910

[35] C. C. Aggarwal, Data Mining: The Textbook. Springer, 2015. https://doi.org/10.1007/978-3-319-14142-8

[36] A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020. https://doi.org/10.1016/j.inffus.2019.12.012

[37] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing design practices for explainable AI user experiences," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-15, 2020. https://doi.org/10.1145/3313831.3376590

[38] M. L. Cummings, "Automation bias in intelligent time critical decision support systems," AIAA 1st Intelligent Systems Technical Conference, 2004. https://doi.org/10.2514/6.2004-6313

[39] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," Nature Machine Intelligence, vol. 1, pp. 501-507, 2019. https://doi.org/10.1038/s42256-019-0114-4