Communication

Shadow AI in Organisations: A Practical Framework for Detection, Risk Classification, and Governance

Ayokunle Ojowa * , Monteiro Marques , Antonio Goncalves

Posted Date: 28 February 2026

doi: 10.20944/preprints202602.1924.v1

Keywords: shadow AI; shadow IT; AI governance; detection; monitoring; compliance; GDPR; EU AI Act; auditability; evidence artefacts

Communication

Shadow AI in Organisations: A Practical Framework for Detection, Risk Classification, and Governance

Ayokunle Ojowa [1,*], Monteiro Marques [2] and Antonio Goncalves [1]

[1] Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
[2] Portuguese Naval Academy / CINAV, 2810-001 Almada, Portugal; mario.monteiro.marques@marinha.pt
[3] Portuguese Naval Academy / CINAV, 2810-001 Almada, Portugal; antonio.leonardo.goncalves@marinha.pt
* Correspondence: ayokunleojowa@tecnico.ulisboa.pt

Abstract

Shadow AI—the unsanctioned use of artificial intelligence tools, models, or services within organisational processes—introduces governance, security, and privacy risks that extend beyond traditional shadow IT. This communication proposes a practical framework to (i) define and classify shadow AI use cases, (ii) detect shadow AI activity through multi-layer technical signals, and (iii) govern risk through an obligations-to-evidence mapping that supports compliance and auditability. The framework aims to balance innovation and productivity with proportionate controls, offering clear remediation paths (block, replace, or regularise with evidence). We also outline a validation plan based on a PRISMA-informed literature review and triangulation (expert feedback, case studies, and survey) to support subsequent empirical evaluation.

Keywords: shadow AI; shadow IT; AI governance; detection; monitoring; compliance; GDPR; EU AI Act; auditability; evidence artefacts

1. Introduction

Artificial intelligence (AI) tools have become widely available and easy to use, which has accelerated their adoption inside organisations. Alongside approved enterprise deployments, many employees now use external AI services, plug-ins, or personal accounts to support daily work tasks. When this happens outside formal approval, risk assessment, and monitoring processes, it creates what we refer to as *shadow AI*: unsanctioned AI use embedded in organisational workflows (IBM, 2024; Palo Alto Networks, 2025).

Shadow AI goes beyond the traditional concept of shadow IT, which primarily involves the unauthorised use of software or IT solutions. In addition to bypassing security controls, shadow AI can introduce new risks related to data leakage (including personal data), unpredictable model behaviour, third-party processing, and hidden decision influence. These risks are difficult to manage because the organisation often lacks visibility into what tools are being used, what data is being shared, and how AI outputs are shaping decisions and actions (Varonis,2025; Wiz, 2025; Zylo,2025).

While existing guidance covers AI governance at a high level, organisations still lack an end-to-end, operational and auditable approach that connects (i) a clear definition and classification of shadow AI, (ii) feasible detection signals across technical layers, and (iii) governance mechanisms that translate regulatory and policy duties into concrete evidence artefacts for audits and continuous oversight. This gap is especially relevant under risk-based regulatory expectations, where accountability, traceability, and appropriate human oversight must be demonstrable (European Parliament, 2024).

To address this need, this communication proposes a practical framework for shadow AI detection and governance. The framework balances productivity and innovation with proportionate controls and provides explicit remediation paths: *block* high-risk uses, *replace* them with approved alternatives, or *regularize* them through governed onboarding supported by evidence. The remainder of this paper

 doi:10.20944/preprints202602.1924.v1

is organized as follows: Section 2 summarizes related work and context; Section 3 defines scope and risk classification; Section 4 presents the proposed framework; Section 5 outlines the validation plan; Section 6 discusses practical and privacy constraints; and Section 7 concludes and highlights future work.

1.1. Contributions

This communication provides:

- A usable definition of shadow AI and a compact taxonomy for risk classification.
- A multi-layer detection view with concrete signal/log categories and known limits.
- An obligations-to-evidence mapping concept to operationalise governance and audit readiness.
- A validation plan (PRISMA-informed review + triangulation) for future empirical assessment.

2. Related Work and Context

2.1. From Shadow IT to Shadow AI

Shadow IT has been extensively studied in information systems research over the past decade. Haag and Eckhardt (2017) define shadow IT as information technology resources used within organisations without explicit organisational approval, encompassing hardware, software, and services that bypass formal IT governance. Systematic literature reviews by Raković et al. (2020) and Klotz et al. (2019) have synthesised research on shadow IT's causes, consequences, and governance approaches, identifying employee dissatisfaction with sanctioned solutions and productivity pressures as primary drivers of unauthorised technology adoption. Silic and Back (2014) characterise shadow IT as a response to perceived gaps between organisational IT provisions and employee workflow requirements.

While shadow AI shares foundational characteristics with shadow IT — namely, unsanctioned adoption driven by productivity goals — it introduces distinct risk dimensions that extend beyond traditional IT governance concerns. Silic et al. (2025) empirically investigated shadow AI as a socio-technical governance failure, demonstrating through survey and interview data that AI's generative, opaque, and autonomous nature creates novel challenges related to data privacy, algorithmic bias, hallucination, and governance drift. Their findings reveal a "governance drift zone" where formal policies exist but lack practical enforcement.

Five key differentiators distinguish shadow AI from its predecessor:

1. **AI Supply Chain Complexity**: Unlike conventional software with deterministic behaviour, AI systems depend on complex supply chains involving third-party models, training data of uncertain provenance, and external inference services (NIST, 2023). When employees use external generative AI services, organisations lose visibility into model versioning, training data composition, and processing pipelines — creating compliance blind spots that do not exist with traditional shadow IT.
2. **Data Leakage Amplification**: Shadow AI magnifies data exposure risks because generative AI tools often process, store, or incorporate user inputs into model training (IBM, 2024). Employees who submit confidential documents, customer data, or proprietary information to external AI services may inadvertently expose sensitive organisational assets to third-party processors without contractual safeguards or data processing agreements.
3. **Decision Impact and Accountability**: AI-generated outputs increasingly influence organisational decisions, from document drafting to data analysis and strategic recommendations. When these outputs originate from unsanctioned tools, the organisation cannot ensure quality, accuracy, or alignment with organisational policies (Puthal, 2025). This creates accountability gaps, particularly in regulated functions such as legal, finance, and human resources.
4. **Opacity and Explainability Deficits**: Shadow AI compounds the inherent interpretability challenges of machine learning systems. Organisations using sanctioned AI can implement monitoring, logging, and explainability mechanisms; shadow AI, by definition, operates outside such

 doi:10.20944/preprints202602.1924.v1

controls. This opacity conflicts with regulatory expectations for transparency and traceability under frameworks such as the EU AI Act (EU AI Act, 2024).

5. **Automation of Influence**: Shadow IT primarily concerns infrastructure and data access; shadow AI extends to automated content generation and decision support. AI-generated text, analysis, or recommendations may propagate through organisational processes without human verification, amplifying the consequences of errors, biases, or inappropriate outputs (CSA, 2024).

The distinction between shadow IT and shadow AI extends beyond technological differences, it encompasses regulatory and accountability dimensions. While shadow IT primarily implicates security and operational risks, shadow AI intersects with fundamental rights considerations (through potential discriminatory outputs), data protection obligations (through training data and prompt processing), and emerging AI-specific regulatory requirements. This regulatory complexity amplifies the governance challenge, as organisations must simultaneously address cybersecurity, privacy, and AI-specific compliance obligations.

2.2. Governance and Regulatory Drivers

The proliferation of shadow AI within organisations occurs against the backdrop of intensifying regulatory scrutiny over AI governance. Two regulatory frameworks are particularly consequential for organisations operating in or serving European markets: the General Data Protection Regulation (GDPR) and the EU Artificial Intelligence Act (EU AI Act). Both frameworks impose obligations that presume organisational visibility and control over technology deployments. These conditions are clearly undermined by the existence of shadow AI. Other relevant frameworks include the NIST AI Risk Management Framework and the ISO/IEC 42001:2023.

1. **GDPR: Data Protection and Automated Decision-Making**
   The GDPR establishes foundational requirements for the lawful processing of personal data, with several provisions bearing directly on AI governance. Article 5 mandates that personal data be processed lawfully, fairly, and transparently, with controllers maintaining accountability for demonstrating compliance (GDPR, 2016). When employees route personal data through unsanctioned AI services, organisations lose the capacity to ensure or demonstrate adherence to these principles.

   Article 22 imposes specific constraints on automated individual decision-making. Data subjects hold the right not to be subject to decisions based solely on automated processing, including profiling, that produce legal effects or similarly significant impacts (Art22 GDPR). Where such processing occurs under permitted exceptions (contractual necessity, legal authorisation, or explicit consent), controllers must implement suitable safeguards, including the right to obtain human intervention, express a viewpoint, and contest the decision. Shadow AI fundamentally compromises these obligations: organisations cannot ensure human oversight of automated decisions when they lack awareness that such decisions are being made, nor can they provide contestability mechanisms for processes invisible to governance structures.

   The transparency obligations under Articles 13 and 14 require controllers to inform data subjects about the existence of automated decision-making and provide meaningful information about the logic involved, as well as the significance and envisaged consequences of such processing (GDPR, 2016). Fulfilling these disclosure requirements presupposes that organisations maintain inventories of AI systems processing personal data.

   Article 35 mandates Data Protection Impact Assessments (DPIAs) for processing operations likely to result in high risk to individuals' rights and freedoms, explicitly including systematic and extensive evaluation of personal aspects based on automated processing (GDPR, 2016). The European Data Protection Board has emphasised that AI systems processing personal data

 doi:10.20944/preprints202602.1924.v1

frequently trigger DPIA requirements (EDPB, 2024). Shadow AI circumvents these safeguards entirely, as organisations cannot assess risks associated with systems they do not know exist.

The accountability principle under Article 5(2) requires controllers to demonstrate compliance with data protection principles (GDPR, 2016). This demonstrable accountability is incompatible with unmonitored AI usage. Organisations cannot produce evidence of lawful processing, appropriate safeguards, or risk mitigation for systems operating outside their governance frameworks.

2. EU AI Act: Risk-Based Governance Obligations

The EU AI Act, which entered into force on 1 August 2024 with phased compliance deadlines extending through 2027, establishes the world's first comprehensive regulatory framework for AI systems. The Act adopts a risk-based approach, calibrating obligations according to the potential harm AI systems may cause to health, safety, or fundamental rights (EU AI Act, 2024).

2.1 **Transparency as a Foundational Requirement**: Transparency is a core pillar of the EU AI Act. Article 13 requires high-risk AI systems to provide sufficient transparency for appropriate interpretation and use. Article 50 extends transparency obligations to certain AI systems regardless of risk classification, including disclosure of AI interaction and machine-readable labelling of AI-generated content (Arts 13, 50 EUAIA). These requirements presuppose organisational awareness of AI deployments; shadow AI, by definition, bypasses such transparency mechanisms.

2.2 **Accountability Through Documentation and Logging**: The EU AI Act embeds accountability through mandatory documentation and logging. Article 11 requires technical documentation for high-risk AI systems, Articles 12 and 19 mandate automatic event logging and log retention, and Article 26(6) obliges deployers to retain logs for at least six months (EUAIA, 2024). Shadow AI operates outside these accountability structures, preventing organisations from producing required documentation or auditable compliance evidence.

2.3 **Deployer Obligations and Human Oversight**: Article 26 imposes specific obligations on deployers of high-risk AI systems, including ensuring compliant use, assigning competent human oversight, maintaining input data quality, and monitoring system operation (Art 26 EUAIA). Article 14 further requires systems to enable effective human oversight by design. Shadow AI fundamentally conflicts with these obligations, as organisations cannot monitor, oversee, or govern systems operating beyond institutional awareness.

2.4 **Governance Implications of Non-Compliance**: The EU AI Act establishes significant penalties for non-compliance, including fines of up to €35 million or 7% of global annual turnover for prohibited practices, and up to €15 million or 3% for other infringements (Art 99 EUAIA). Beyond financial sanctions, non-compliance carries reputational and market-access risks. Shadow AI creates compliance blind spots that expose organisations to enforcement actions for obligations they may be unaware they are breaching.

3. Complementary Governance Frameworks

Beyond mandatory compliance, voluntary governance frameworks offer practical implementation guidance that strengthens the governance objectives set by regulation. The NIST AI Risk Management Framework (AI RMF 1.0), released in January 2023, offers a structured approach to AI risk identification, assessment, and mitigation. It is organised around four core functions: Govern, Map, Measure, and Manage (NIST 2023). The framework emphasises transparency, accountability, and explainability, all of which assumes organisational awareness and control of AI systems.

ISO/IEC 42001:2023 establishes requirements for AI Management Systems (AIMS), providing a certifiable standard for organisational AI governance (ISO 42001). The standard addresses AI-specific management challenges, including transparency, explainability, and continuous learning

 doi:10.20944/preprints202602.1924.v1

behaviours that require special consideration for responsible use (ISO 42001). Implementation of ISO/IEC 42001 necessarily requires comprehensive inventories of AI systems within the organisation's scope. It is expected that such inventories should not account for shadow AI.

3. Shadow AI: Operational Scope and Risk Lens

Effective governance requires precise scoping. This section establishes an operational definition of shadow AI, delineates boundary conditions to distinguish in-scope from out-of-scope phenomena, and introduces a multi-dimensional taxonomy for risk classification that enables proportionate governance responses.

3.1. Operational Definition and Boundaries

We define **shadow AI** as the use of artificial intelligence tools, models, or services within organisational workflows without formal authorisation, risk assessment, or governance oversight. This definition extends the established concept of shadow IT (Haag, 2017; Silic, 2014) to encompass AI-specific characteristics: probabilistic outputs, opaque processing chains, third-party inference services, and the potential for automated decision influence.

Three criteria jointly characterise shadow AI: (1) *unsanctioned status* (the AI tool or service has not received formal organizational approval); (2) *work-related use* (the tool is employed for tasks within the scope of employment or organizational function); and (3) *governance bypass* (usage occurs outside established risk assessment, security review, or compliance processes).

To operationalise this definition, explicit boundary conditions are defined. The following examples present representative in-scope and out-of-scope scenarios to support consistent detection and classification.

Includes (In-scope)

The following use cases fall within the operational scope of shadow AI:

- Use of public generative AI services (e.g., ChatGPT, Claude, Gemini) via browser or mobile application for work tasks without organisational approval.
- Installation of unapproved browser extensions, plug-ins, or desktop applications that transmit organizational content to third-party AI services for processing.
- Direct API calls to external AI inference endpoints using personal accounts or credentials, where outputs are incorporated into organisational deliverables.
- Embedding of AI-powered features within personal SaaS subscriptions (e.g., AI writing assistants, code completion tools) used for work purposes.
- Local deployment of open-source large language models on personal devices for work-related tasks without IT security review.
- Use of AI features auto-enabled within sanctioned applications where the AI component itself has not undergone separate governance review.

Excludes (Out-of-scope)

The following scenarios are explicitly out of scope:

- Enterprise AI features embedded within sanctioned SaaS platforms operating under contractual data processing agreements and organisational governance.
- Internal AI systems deployed through official MLOps pipelines with documented risk assessment, security review, and compliance sign-off.
- AI tools procured through formal vendor evaluation processes with established service-level agreements and data protection addenda.
- Personal use of AI tools for non-work purposes on personal time and personal devices, with no organisational data involved.

 doi:10.20944/preprints202602.1924.v1

- Research and development use within designated sandbox environments with appropriate access controls and data segregation.

3.2. Taxonomy for Risk Classification

Shadow AI risk is not uniform; governance responses must be calibrated to the nature and severity of potential harms. A five-dimensional taxonomy is proposed to enable systematic risk classification, drawing on established frameworks including the EU AI Act's risk-based approach (EUAIA, 2024), NIST AI RMF's trustworthiness characteristics (NIST, 2023), and ISO/IEC 42001's AI management system requirements (ISO 42001).

3.3. Risk Dimensions

Each shadow AI instance can be characterised along five independent dimensions. Together, these dimensions capture the principal factors influencing potential harm and regulatory exposure, enabling systematic risk scoring.

D1: Data Sensitivity

The classification of data processed through shadow AI channels determines exposure severity. Categories include:

- *Public*: Non-confidential, publicly available information.
- *Internal*: Organisational information not intended for external distribution.
- *Confidential*: Proprietary business information, trade secrets, strategic plans.
- *Regulated*: Personal data subject to GDPR (GDPR, 2016), health information, financial records, or other legally protected categories.

D2: Decision Impact

The degree to which AI outputs influence consequential decisions determines accountability requirements. Levels include:

- *Informational*: AI output used for background research or ideation with no direct decision influence.
- *Advisory*: AI output informs human decisions but undergoes substantive review before action.
- *Determinative*: AI output materially shapes decisions affecting individuals, contracts, or organisational commitments.
- *Automated*: AI output directly triggers actions with legal or similarly significant effects, potentially engaging Article 22 GDPR obligations (Art22, GDPR).

D3: Processing Transparency

The visibility into AI processing chains affects auditability and compliance demonstration:

- *Documented*: Processing logic, data flows, and model provenance are known and recorded.
- *Partially Opaque*: Service provider discloses general processing approach but not implementation details.
- *Fully Opaque*: No visibility into model architecture, training data, or processing pipeline; "black box" third-party service.

D4: Organisational Function

Certain functional domains carry elevated regulatory scrutiny or reputational sensitivity:

- *General Operations*: Administrative, logistical, or support functions.
- *Customer-Facing*: Marketing, sales, customer service, where AI outputs may reach external parties.
- *High-Stakes Functions*: Human resources, legal, finance, compliance – domains with significant individual impact or regulatory exposure.

 doi:10.20944/preprints202602.1924.v1

- *Safety-Critical*: Functions where errors could cause physical harm, system failures, or significant financial loss.

D5: Third-Party Dependency

The nature of external AI service relationships affects data protection and continuity risks:

- *None*: Locally deployed models with no external data transmission.
- *Inference-Only*: Data transmitted to external service for processing; no declared retention or training use.
- *Training-Inclusive*: Service terms permit use of inputs for model improvement, creating data leakage and IP exposure.
- *Unknown*: Terms of service not reviewed or understood; data handling practices indeterminate.

3.4. Risk Tier Assignment

The five dimensions combine to produce an aggregate risk tier that guides governance response. Table 1 presents the tiered classification schema.

Table 1. Shadow AI Risk Tiers and Governance Implications.

Tier	Characterization	Governance Response
Critical	Regulated/confidential data + determinative/automated decisions + high-stakes functions	Immediate block; incident investigation; remediation mandatory
High	Confidential data OR determinative decisions OR high-stakes functions	Replace with approved alternative; escalate to governance review
Moderate	Internal data + advisory decisions + partially opaque processing	Regularize through governed onboarding with evidence artefacts
Low	Public data + informational use + general operations	Monitor; consider for approved tool catalogue; awareness training

This tiered approach enables proportionate governance: critical-tier shadow AI warrants immediate intervention, while low-tier instances may be candidates for regularisation into approved usage with appropriate controls. The taxonomy thus supports the framework's remediation pathways (block, replace, regularise) detailed in Section 4.

4. Proposed Framework

This section presents the core contribution: an integrated framework for shadow AI detection and governance. The framework comprises two complementary components: (1) a lifecycle model that structures governance activities from initial discovery through continuous improvement, and (2) an obligations-to-evidence mapping that operationalises regulatory requirements into auditable artefacts. Together, these components enable organisations to establish visibility over shadow AI usage and demonstrate compliance with applicable regulatory frameworks.

4.1. Lifecycle: From Discovery to Continuous Control

Effective shadow AI governance requires a structured, iterative approach rather than one-time intervention. A six-stage lifecycle model that guides organisations from initial detection through sustained oversight is proposed. Each stage produces defined outputs that feed subsequent stages, creating a closed-loop governance process aligned with continuous improvement principles embedded in ISO/IEC 42001 (ISO 42001) and the NIST AI RMF's iterative risk management approach (NIST, 2023).

Stage 1: Discover

The discovery stage establishes visibility over AI tool usage within the organisation. Detection mechanisms operate across multiple technical layers (detailed in Section 4.4), including network-level

 doi:10.20944/preprints202602.1924.v1

signals, endpoint telemetry, identity and access logs, and behavioural analytics. The output of this stage is an inventory of detected shadow AI instances, each characterised by tool identifier, user or department, access pattern, and initial data exposure indicators.

Stage 2: Assess

Discovered instances undergo systematic risk assessment using the five-dimensional taxonomy presented in Section 3.2. Assessment considers data sensitivity, decision impact, processing transparency, organisational function, and third-party dependency. Each instance receives a risk tier assignment (Critical, High, Moderate, or Low) that determines governance priority and remediation pathway. Assessment should also identify potential regulatory implications, particularly where personal data processing may require GDPR obligations (GDPR, 2016) or where use cases align with EU AI Act high-risk categories (EUAIA, 2024).

Stage 3: Decide

Based on risk tier and organisational context, governance stakeholders determine the appropriate remediation pathway for each shadow AI instance:

- **Block**: Prohibit and technically restrict usage. Appropriate for Critical-tier instances involving regulated data, prohibited AI practices, or unacceptable risk exposure.
- **Replace**: Substitute with an approved alternative that meets security, privacy, and compliance requirements while preserving productivity benefits.
- **Regularise**: Onboard the tool or use case into governed AI operations through formal risk assessment, contractual arrangements, and evidence artefact generation.

Decision criteria should be documented and consistently applied, with escalation pathways for edge cases requiring senior governance review.

Stage 4: Remediate

Remediation executes the decision through technical controls, policy enforcement, and user engagement. For blocked instances, this includes network-level restrictions, endpoint controls, and user notification. For replacement, this involves provisioning approved alternatives and supporting user transition. For regularisation, remediation encompasses vendor due diligence, data processing agreement negotiation, integration with organisational AI governance processes, and generation of required evidence artefacts per the obligations-to-evidence mapping (Section 4.2).

Stage 5: Monitor

Ongoing monitoring ensures remediation effectiveness and detects new shadow AI emergence. Monitoring encompasses both technical surveillance (continuous detection across network, endpoint, and identity layers) and governance oversight (periodic review of regularised tools, audit of evidence artefact currency, and verification of control effectiveness). Monitoring outputs feed back into the Discover stage, creating continuous visibility. This aligns with EU AI Act Article 26 requirements for deployers to monitor high-risk AI system operation (Art26 EUAIA) and GDPR accountability obligations requiring demonstrable compliance (GDPR, 2016).

Stage 6: Improve

Governance processes undergo periodic review and refinement based on accumulated experience. Improvement activities include updating detection signatures for emerging AI tools, refining risk assessment criteria based on incident patterns, streamlining regularisation pathways to reduce friction, and enhancing user awareness programs. Lessons learned from shadow AI incidents inform organisational AI strategy and policy development.

Figure 1 illustrates the lifecycle stages and their interconnections.

 doi:10.20944/preprints202602.1924.v1

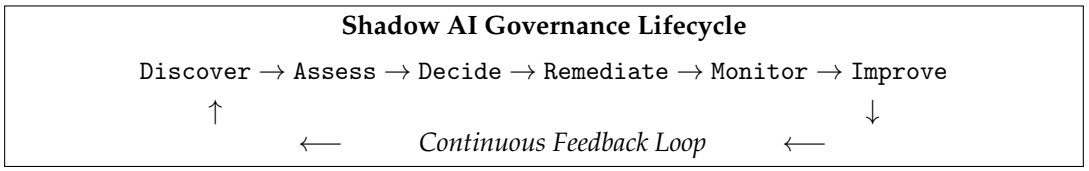

Figure 1. Six-stage lifecycle for shadow AI governance, with continuous feedback from monitoring and improvement stages to discovery.

4.2. Obligations-to-Evidence Mapping

A central challenge in AI governance is translating abstract regulatory requirements into concrete organisational practices. The obligations-to-evidence mapping addresses this challenge by establishing explicit linkages between regulatory duties and verifiable artefacts that demonstrate compliance. This approach operationalises the accountability principle central to both GDPR Article 5(2) (GDPR, 2016) and EU AI Act governance requirements (EUAIA, 2024).

4.3. Mapping Structure

The mapping follows a structured template that decomposes regulatory requirements into actionable governance elements. Table 2 presents the mapping schema.

Table 2. Obligations-to-Evidence Mapping Structure.

Element	What it captures
Requirement	The regulatory or policy provision that imposes a verifiable duty (e.g., GDPR Article 35, EU AI Act Article 26).
Obligation	The operational statement of what must be ensured, expressed in terms that are open to verification.
Evidence Artefact	The specific document, log, record, or report that demonstrates obligation fulfilment.
Control	The technical and/or organisational mechanism that produces or supports the evidence artefact.
Owner (RACI)	Assignment of Responsible, Accountable, Consulted, and Informed roles for the obligation.
Trigger	The condition or event that initiates review, escalation, or remediation action.

4.4. Multi-Layer Detection View

Detection is the prerequisite for governance, i.e., organisations cannot manage shadow AI they cannot see. This section presents a multi-layer detection architecture that combines signals across network, endpoint, identity, behavioural, and content inspection layers. No single layer provides complete visibility, so effective detection requires triangulation across complementary data sources, each with its own strengths and limitations.

 doi:10.20944/preprints202602.1924.v1

Table 3. Detection layers, example signals, and key limitations (high-level).

Layer	Observable signals / log sources	Main limitations
Network	DNS/HTTP(S) metadata, proxy logs, SASE logs, CASB signals.	Encryption, personal networks, approved SaaS overlap.
Endpoint	EDR telemetry, process/browser events, clipboard/file activity patterns.	BYOD, privacy constraints, local models.
Identity/SaaS	SSO audit logs, OAuth grants, unusual app authorisations.	Shadow accounts, token reuse, incomplete coverage.
UBA/Behaviour	Anomalous usage patterns, time/volume irregularities.	False positives, baseline drift.
Content/DLP (where applicable)	Policy matches for sensitive data exfiltration.	Minimisation limits, encryption, scope constraints.

4.5. Governance and Remediation Paths

Detection and assessment produce actionable intelligence, which are translated via governance into organisational response. The framework adopts a risk-proportionate approach: remediation severity scales with assessed risk tier, balancing security and compliance needs against productivity and innovation value. This calibration reflects the EU AI Act's risk-based philosophy (EUAIA, 2024) and avoids the governance failure of uniform prohibition, which historically drives shadow IT deeper underground rather than eliminating it (Klotz, 2019).

Decision Logic

Remediation pathway selection follows risk tier assignment (Section 3.2) and organisational context. Figure 2 illustrates the decision flow.

Risk Tier → Primary Pathway

Critical: Regulated data + automated decisions + high-stakes functions → **Block**
High: Confidential data OR determinative impact OR sensitive functions → **Block** or **Replace**
Moderate: Internal data + advisory use + partial opacity → **Replace** or **Regularise**
Low: Public data + informational use + general operations → **Regularise** or **Monitor**

Contextual factors: Business criticality, alternative availability, user population, remediation cost

Figure 2. Risk-tier-to-remediation pathway mapping with contextual modifiers.

Remediation Options

- **Block**: prohibit and technically restrict high-risk shadow AI usage.
- **Replace**: provide an approved alternative that meets security/compliance needs.
- **Regularise**: onboard the use case into governed AI with explicit evidence artefacts.

5. Validation Plan

This communication presents a conceptual framework requiring empirical validation to establish practical utility and refine the proposed constructs. We adopt a mixed-methods validation strategy that combines evidence synthesis, expert feedback, case studies, and survey. This triangulated approach addresses different validity dimensions: theoretical grounding through literature review, construct validity through expert feedback/assessment, practical applicability through case studies, and perceived usefulness through survey instrumentation.

We will evaluate the framework through:

- **PRISMA-informed systematic review**: A structured literature review following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page, 2021)

 doi:10.20944/preprints202602.1924.v1

will identify and synthesise existing evidence on shadow AI risks, detection mechanisms, and governance approaches.

- **Expert feedback**: Reliance on domain experts to assess framework completeness, feasibility, and clarity of the mapping and detection layers. The expert panel will consist of practitioners from information security, data protection, AI governance, and compliance functions across multiple industry sectors.
- **Case studies**: The framework will be applied to representative organisational scenarios to assess practical applicability. Case study design follows established information systems research methodology (Yin, 2018).
- **Survey**: A Survey instrument will be used to measure perceived usefulness, adoption barriers, and governance acceptability among IT security, compliance, and business professionals.

6. Discussion

The proposed framework addresses a genuine governance gap, but implementation can be affected by significant constraints. This section examines three critical considerations: the tension between detection and workplace privacy, practical limitations affecting framework effectiveness, and forward-looking extensions that may strengthen governance as the regulatory and technical landscape evolves.

6.1. Privacy-Preserving Monitoring in the Workplace

Shadow AI detection involves workplace monitoring, which increases data protection concerns that must be addressed for lawful and ethical implementation. The framework's detection mechanisms process employee data and are subject to regulatory requirements and applicable national legislations.

Five principles guide privacy-preserving implementation: ***proportionality***, ***data minimisation***, ***retention limits***, ***transparency***, and ***impact assessment***.

- **Proportionality**: Monitoring must be limited to what is necessary for security objectives; targeted, risk-based detection is preferable to broad surveillance and should be supported by a documented proportionality assessment.
- **Data Minimisation**: Detection should rely on the least intrusive data possible, prioritising metadata over content inspection, with human review restricted to confirmed policy violations.
- **Retention Limits**: Monitoring data or detection logs should be retained only for predefined, justified periods, enforced through automated deletion and transparently documented.
- **Transparency**: Employees must be clearly informed about monitoring activities, their purposes, and data rights. Covert monitoring must be avoided except where legally and ethically justified.
- **DPIA Requirement**: Employee monitoring typically requires a GDPR Data Protection Impact Assessment under GDPR Article 35 (GDPR, 2016) to assess privacy risks, proportionality, and mitigations. Employee representatives should be consulted where required by national law.

6.2. Limitations and Practical Constraints

The framework operates within technical, organizational, and methodological constraints that tend to limit its effectiveness.

Technical Limitations

- *Encryption and TLS*: Because most network traffic is encrypted, monitoring is usually limited to basic metadata. Inspecting actual content requires endpoint tools or breaking encryption, which is more complex and raises privacy issues.
- *BYOD and Personal Devices*: Employees accessing AI services via personal devices on personal networks entirely bypass corporate detection infrastructure. Policy-based controls (acceptable use agreements) substitute for technical detection in these scenarios.

 doi:10.20944/preprints202602.1924.v1

- *Shadow Channels*: Browser-based AI access, mobile applications, and API integrations create multiple access vectors. Therefore, comprehensive detection requires coverage across all channels, which may be technically or economically infeasible.
- *Local Models*: Locally deployed open-source models (e.g., running LLaMA variants on personal hardware) generate no network signals to external services, rendering network-based detection ineffective.
- *Approved SaaS with Embedded AI*: Distinguishing shadow AI from AI features embedded in approved SaaS platforms requires maintained inventories and distinct policy definitions.
- *Detection Accuracy*: False positives increase operational burden and reduce trust, while false negatives allow shadow AI usage to go undetected. Detection accuracy is challenged by evolving AI tools and usage patterns, requiring continuous measurement and acceptance of inherent limitations.

Organisational Factors

Framework effectiveness depends on organisational enablers beyond technical controls. Such as governance authority and management buy-in; cross-functional coordination among IT security, legal, compliance, and business units; resource availability; and organisational culture.

Methodological Constraints

The framework is conceptual and has not yet been empirically validated. Its applicability across different organisations, industries, and legal contexts is still uncertain, and ongoing changes in AI tools may require regular updates.

6.3. Regulatory Change and XAI-as-Evidence

The framework can be strengthened by stress-testing obligations-to-evidence mappings against evolving regulations, including GDPR, the EU AI Act, and sector-specific rules. This can help uncover gaps and maintain governance agility. Explainable AI (XAI) outputs, such as feature attributions, counterfactuals, model cards, and audit logs can serve as auditable evidence supporting transparency under EU AI Act Article 13 and GDPR Articles 13–14. Integrating XAI positions explainability as a governance tool, enabling responsible AI deployment to be clearly documented.

7. Conclusions

This communication examines the emerging organisational risk associated with *shadow AI*—artificial intelligence systems and tools that are developed, deployed, or used without formal authorisation, security review, or governance oversight. Extending the traditional *shadow IT* problem, shadow AI introduces additional operational and regulatory complexity due to probabilistic outputs, opaque processing and integration chains, and increased exposure to data protection and compliance obligations, notably under the European Union Artificial Intelligence Act (EU AI Act) and the General Data Protection Regulation (GDPR).

The principal conclusion is that effective shadow AI management requires an integrated *techno-legal* response. In practice, organisations must combine (i) legal and regulatory mapping that translates obligations into actionable governance artefacts, with (ii) technical detection and monitoring capabilities that provide defensible visibility into unauthorised AI usage, while remaining proportionate and privacy-preserving. Consistent with the proposed work plan, the methodological foundation is anchored in a systematic literature review using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology, complemented by a solution-oriented research design and a multi-method evaluation strategy (expert validation, case studies, and surveys). This combination is intended to ensure that the proposed outputs can be substantiated, assessed, and iteratively improved in a way that is both academically rigorous and organisationally implementable.

A primary outcome is an integrated framework architecture that consolidates four essential dimensions into a coherent governance-and-detection capability:

 doi:10.20944/preprints202602.1924.v1

1. **Conceptual foundations**: definition, taxonomy, and lifecycle of shadow AI;
2. **Compliance mechanisms**: regulatory matrices, liability allocation logic, and contractual/organisational templates;
3. **Technical detection and monitoring**: a multi-layered detection architecture incorporating privacy-preserving methods;
4. **Governance and risk management**: a structured, tiered governance model supported by risk assessment instruments, policies, and implementation guidance.

Taken together, these components aim to operationalise regulatory requirements into concrete organisational processes and auditable artefacts, enabling risk reduction without unnecessarily constraining legitimate innovation and productivity.

This communication also consolidates the anticipated contributions across three domains:

- **Theoretical contribution**: a structured conceptualisation of shadow AI as distinct from shadow IT, and a unified framing of how legal requirements can inform technical design (and vice versa) within risk-based organisational governance.
- **Methodological contribution**: the application of PRISMA-driven evidence synthesis to an interdisciplinary AI governance problem, combined with a coherent, mixed evaluation approach (expert validation, case studies, and surveys) that explicitly accounts for data protection constraints.
- **Practical contribution**: delivery of ready-to-adopt organisational tools, including phased implementation roadmaps, regulatory compliance matrices and checklists, privacy-preserving detection guidance, and risk scoring matrices and assessment templates for systematic shadow AI risk evaluation.

Finally, the work recognises key challenges likely to influence execution and evaluation, including constraints on access to empirical organisational data, the rapidly evolving landscape of AI tools and usage patterns, interdisciplinary complexity across security, legal, and organisational domains, and stakeholder diversity that can affect adoption and measurement. These constraints motivate an iterative refinement cycle and establish clear next steps for future work: empirical validation of detection effectiveness and trade-offs (including false positives/negatives), feasibility and acceptability assessment in organisational contexts, refinement of governance triggers and escalation pathways, and extension to cross-jurisdictional governance scenarios.

Author Contributions: Conceptualisation, Ayokunle Ojowa; methodology, Ayokunle Ojowa and Antonio Goncalves ; formal analysis, Ayokunle Ojowa; Mario Marques and Antonio Goncalves investigation, Ayokunle Ojowa; resources, Ayokunle Ojowa; writing original draft preparation, Ayokunle Ojowa; writing—review and editing, Ayokunle Ojowa ; Antonio Goncalves and Mario Marques; visualization, Ayokunle Ojowa; supervision, Ayokunle Ojowa and Antonio Goncalves; project administration, Ayokunle Ojowa funding acquisition, Mario Marques. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The case study uses synthetically generated data produced by the described workflow; no personal data were used. Data can be regenerated from the configuration and generation procedures reported in this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. American Bar Association. Formal Opinion 512: Generative Artificial Intelligence Tools, 2023. ABA Standing Committee on Ethics and Professional Responsibility.
2. Cloud Security Alliance. Shadow AI: Risks and Governance Considerations, 2024. Online.
3. Data Protection Commission (Ireland). Guidance on Artificial Intelligence and Data Protection, 2024. Online.

 doi:10.20944/preprints202602.1924.v1

4. European Data Protection Board. Opinion 28/2024 on Data Protection and Artificial Intelligence Models, 2024. Online.
5. Directive (EU) 2019/1937 of the European Parliament and of the Council on the Protection of Persons Who Report Breaches of Union Law (Whistleblower Protection Directive), 2019. EU Directive 2019/1937.
6. 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 (GDPR), 2016. General Data Protection Regulation.
7. European Commission. The AI Act, 2024. Online.
8. European Parliament. Artificial Intelligence Act: MEPs Adopt Landmark Law, 2024. Online.
9. Fürstenau, D.; Rothe, H. Shadow IT Systems: Discerning the Good and the Evil. In Proceedings of the Proceedings of the 22nd European Conference on Information Systems (ECIS), Tel Aviv, Israel, 2014.
10. Gartner. AI Governance Frameworks for Enterprises, 2024. Gartner Research.
11. GDPR Local. GDPR and AI Compliance Guide, 2025. Online.
12. Haag, S.; Eckhardt, A. Shadow IT. *Business & Information Systems Engineering* **2017**, *59*, 469–473. https://doi.org/10.1007/s12599-017-0497-x.
13. Hevner, A.R.; March, S.T.; Park, J.; Ram, S. Design Science in Information Systems Research. *MIS Quarterly* **2004**, *28*, 75–105.
14. IBM. What Is Shadow AI?, 2024. Online.
15. IEEE. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems, 2019. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.
16. Infosecurity Magazine. Survey Reveals 77% of UK Cyber Leaders Believe GenAI Has Contributed to Rise in Security Incidents, 2024. Online.
17. ISACA. Shadow AI Governance: Managing the Risks, 2025. Online.
18. International Organization for Standardization. ISO/IEC 42001:2023 — Information Technology — Artificial Intelligence — Management System, 2023. Standard.
19. King & Spalding. Navigating the Global AI Regulatory Landscape, 2025. Online.
20. Kopper, A.; Westner, M. Deriving a Framework for Causes, Consequences and Governance of Shadow IT from Literature. In Proceedings of the Proceedings of the 22nd Americas Conference on Information Systems (AMCIS), San Diego, CA, 2016.
21. Kuner, C.; Bygrave, L.A.; Docksey, C. *The EU General Data Protection Regulation (GDPR): A Commentary*; Oxford University Press, 2020.
22. Lakera. Shadow AI: The Hidden Risks Organizations Face, 2024. Online.
23. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gøtzsche, P.C.; Ioannidis, J.P.A.; Clarke, M.; Devereaux, P.J.; Kleijnen, J.; Moher, D. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Medicine* **2009**, *6*, e1000100. https://doi.org/10.1371/journal.pmed.1000100.
24. Lohmann, S.; Krcmar, H. Cloud Access Security Brokers: Current Solutions and Challenges. *Business & Information Systems Engineering* **2018**, *60*, 197–210.
25. McKinsey & Company. The State of AI Governance in 2024, 2024. Online/Report.
26. Microsoft.; LinkedIn. 2024 Work Trend Index: Annual Report, 2024. Online.
27. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine* **2009**, *6*, e1000097. https://doi.org/10.1371/journal.pmed.1000097.
28. National Institute of Standards and Technology. AI Risk Management Framework (AI RMF 1.0), 2023. Online.
29. OECD. Recommendation of the Council on Artificial Intelligence, 2019. OECD/LEGAL/0449.
30. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. https://doi.org/10.1136/bmj.n71.
31. Palo Alto Networks. Shadow AI Security: Detecting and Managing Unauthorized AI, 2025. Online.
32. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems* **2007**, *24*, 45–77. https://doi.org/10.2753/MIS0742-1222240302.
33. PurpleSec. Shadow IT Statistics and Security Risks, 2025. Online.

 doi:10.20944/preprints202602.1924.v1

34. Puthal, D.; Mishra, A.K.; Mohanty, S.P.; Others. Shadow AI: Cyber Security Implications, Opportunities and Challenges in the Unseen Frontier. *SN Computer Science* **2025**, *6*, 405. https://doi.org/10.1007/s42979-025-03962-x.
35. PwC. EU AI Act: What Organizations Need to Know, 2024. Online.
36. Sartor, G.; Lagioia, F. The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence, 2020. European Parliamentary Research Service.
37. Securiti. Shadow AI: Understanding the Risks and Governance Challenges, 2024. Online.
38. Silic, M.; Back, A. Shadow IT — A View from Behind the Curtain. *Computers & Security* **2014**, *45*, 274–283. https://doi.org/10.1016/j.cose.2014.06.003.
39. Singh, A.P.; Kumar, R.; Sharma, S.; Kumar, A. Encrypted Malware Detection Methodology Without Decryption Using Deep Learning-Based Approaches. *Turkish Journal of Engineering* **2024**, *8*, 498–509. https://doi.org/10.31127/tuje.1416933.
40. Varonis. Shadow AI Security Risks: What You Need to Know, 2025. Online.
41. World Health Organization. Ethics and Governance of Artificial Intelligence for Health, 2024. Online.
42. Wiz. Shadow AI: The Invisible Security Threat in Your Organization, 2025. Online.
43. Zylo. Shadow AI Report: Understanding the Unauthorized AI Landscape, 2025. Online.