Headline

Shadow AI : Detection, Control, and Governance for Manufacturing Confidentiality

Won-jun Song, SK Shieldus

1. Strategic Security Threats Posed by Shadow AI in Manufacturing Environments

Since 2023, the commercialization LLM-based generative artificial intelligence (Generative AI) has dramatically accelerated innovation across all industries, driving advances in process automation, engineering optimization, and knowledge refinement. In particular, the manufacturing sector has witnessed the rapid emergence of AI's utility in diverse functional domains, including product design, quality management, process control, and productivity enhancement. However, these technological advancements have simultaneously reshaped the security landscape, introducing new vectors of risk—foremost among them is the phenomenon of Shadow AI.

Shadow AI refers to the unauthorized and informal use of AI services by individuals or departmental units without official organizational approval. In knowledge-intensive industries such as manufacturing, the unsanctioned external transmission of critical assets—including trade secrets, production recipes, routing information, design blueprints, and equipment logic—can pose severe security threats. Such practices can circumvent the detection capabilities of existing security infrastructures (such as DLP, EDR, CASB), resulting in a substantial degradation of security visibility within the organization.

For instance, consider scenarios in which R&D engineers describe CAD designs to LLM-based chatbots in order to solicit technical feedback, or cases where manufacturing technology teams input proprietary process data to optimize production recipes. The critical issue here is that most of these inputs are transmitted as unstructured API traffic over HTTPS, thereby evading internal audit and access control mechanisms. Should such prompts be leveraged as training data or stored long-term by external AI services, there exists a tangible risk that proprietary information may be repurposed in subsequent model training within the same industry sector.

Moreover, the risks associated with Shadow AI extend beyond information leakage, encompassing secondary threats such as regulatory non-compliance, legal disputes, and violations of industrial protection statutes. Notably, under domestic and international regulatory frameworks—such as the Industrial Technology Protection Act, GDPR, and ITAR—the mere loss of control over confidential information is sufficient grounds for forfeiting its protected status. As a result, even a single instance of external transmission may irreversibly compromise the legal protection afforded to patent assets.

Accordingly, Shadow AI must not be dismissed as a mere 'user behavior issue'; rather, it should be recognized as a structural vulnerability within knowledge-driven security strategies for the manufacturing sector. This reality underscores the urgent need to establish a comprehensive governance model encompassing proactive detection, behavioral control, and prompt-level risk assessment frameworks.

In this Insight, we examine the operational dynamics of Shadow AI and its manufacturing-specific threat scenarios, and propose an effective security model that encompasses both technical detection mechanisms and policy-driven response strategies. Furthermore, drawing upon global regulatory trends and response guidelines, we present a reference framework designed to support the establishment of practical, operations-oriented governance systems.

2. Conceptual Overview and Threat Model Analysis

2.1 Definition and Behavioral Characteristics of Shadow AI

Shadow AI refers to the practice whereby individuals or departmental units utilize unauthorized generative AI tools—such as large language models (LLMs), Vision AI, or AutoML—without passing through the organization's established security or IT management frameworks. In this process, users often engage in the following behaviors, frequently without adequate awareness of security protocols or data handling regulations.

- Directly inputting internal documents, blueprints, or process information into external AI systems in the form of prompts
- Integrating code or documents generated by external AI into operational systems without proper validation
- Failing to recognize that sensitive data may be automatically stored or cached on external servers outside the corporate perimeter

While the use of Shadow AI may ostensibly aim to enhance workplace productivity and support individual tasks, from a security perspective it constitutes the high-risk transmission of sensitive data through unauthorized channels.

2.2 Shadow AI Threat Model Classification (Manufacturing-Centric)

The following section delineates the various threat types associated with Shadow AI in manufacturing environments, structured around behavior, risk, impact, and illustrative examples.

① Leakage of Design and Technical Documentation

Element	Description
Behavior	Requesting explanations of CAD drawings or summaries of product design structures
Risk	Exposure of design expertise, component specifications, and positioning information to LLMs
Impact	Potential exploitation for imitation of similar products or acquisition of proprietary technology by competitors
Example	Including the complete design structure in a prompt such as, “Is there any overall issue with this design?”

② Exposure of Manufacturing Recipes and Process Parameters

Element	Description
Behavior	Querying AI for process condition adjustments or methods to improve yield
Risk	Transmission of internal variables such as production temperature, speed, and material ratios
Impact	Loss of quality competitiveness; transfer of proprietary information to OEM/ODM competitors
Example	Prompting with questions like, “Analyze the causes of defects for this material ratio.” thereby disclosing sensitive process details

③ Leakage of Sensitive Information via Quality Data

Element	Description
Behavior	Inputting defect occurrence databases, inspection images, or defect types into AI systems
Risk	Product defect data and structural vulnerability information are learned by external entities
Impact	Potential identification of vulnerable products, which could be exploited to maliciously trigger recalls
Example	Prompts such as, “Explain why this photo was classified as a grade B defect.” inadvertently disclose sensitive quality data

④ Leakage and Compromise of Automation Code or Sequences

Element	Description
Behavior	Requesting AI to diagnose PLC control code or sequence logic
Risk	Exposure of code logic, or incorporation of insecure logic from AI-generated code
Impact	Potential for equipment shutdown, safety incidents, or propagation of attacks targeting operational technology (OT) systems
Example	AI-generated code omits authentication procedures, enabling injection of external commands

⑤ Indirect Leakage of User Credentials and System Information

Element	Description
Behavior	Supplying LLMs with development code or API examples
Risk	Disclosure of authentication tokens, account names, and system port configurations
Impact	Unintentionally furnishing attackers with a blueprint of internal APIs
Example	Requests such as, “Show me how to integrate this API with the quality management system.” which may inadvertently reveal sensitive system architecture details

⑥ Information Inference via Training Data Reuse

Element	Description
Behavior	Repeatedly inputting prompts containing internal information into LLMs
Risk	Subsequent prompts from other users may elicit generated responses that reproduce the previously entered confidential data
Impact	Loss of confidentiality, effectively equivalent to public disclosure of the information
Example	Requests such as, “Show me the production recipe I provided earlier.” resulting in sensitive data being resurfaced in model outputs

⑦ Regulatory and Compliance Violations

Element	Description
Behavior	Transmitting confidential information to overseas AI servers, potentially violating regulations such as GDPR, ITAR, or the Industrial Technology Protection Act
Risk	Non-compliance with regulatory requirements, exposure to legal action, and risk of certification revocation
Impact	Damage to corporate reputation and loss of external contracts
Example	Transmission of design blueprints from a defense component manufacturer to OpenAI

As demonstrated by the aforementioned cases, Shadow AI exhibits the following multifaceted characteristics:

- Low-intent, High-impact : While user actions may be well-intentioned, their consequences can prove catastrophic.
- Technical Undetectability : Information embedded within prompts is inherently difficult to identify and classify using conventional methods.
- Governance Externality : Such activities occur outside the purview of traditional information security management frameworks.
- Expansion of the Attack Surface: External API and model invocations effectively create new security perimeters.

2.3 Derivation of Key Issues

Within manufacturing organizations, Shadow AI should not be dismissed as mere employee negligence; rather, it constitutes a warning sign that exposes fundamental deficiencies in the organization's security governance framework. Even in the absence of an external attacker, critical assets can be exfiltrated internally, and any leaked information remains irretrievable—necessitating that such incidents be classified as irreversible security breaches.

Accordingly, the detection, prevention, mitigation, and incident response for Shadow AI must be regarded not as optional measures, but as indispensable elements of security strategy in the era of digital manufacturing.

3. Technical Response Strategies: Detection, Control, and Mitigation

3.1 Detection Strategy

Visibility is paramount for the effective detection of Shadow AI. To accurately identify HTTPS-based AI API calls, dynamic domains, and unstructured prompts, the following response framework is recommended.

① Leakage of Design and Technical Documentation

- AI platform calls can be identified through Server Name Indication (SNI), User-Agent, and Domain Name System (DNS) request patterns
- Advanced Cloud Access Security Broker (CASB) solutions enable real-time detection and policy enforcement for external Large Language Model (LLM) API calls
- However, conventional CASB platforms provide limited detection capabilities for Shadow AI; therefore, Data Security Posture Management (DSPM) features capable of capturing AI-related data flows are required

② Prompt Content-Based Anomaly Detection

- Implement policies to detect high-risk keywords such as “design,” “confidential,” “process,” or “revenue,” flagging prompts that contain sensitive information
- Apply AI-aware Data Loss Prevention (DLP) mechanisms or prompt injection detection rules to monitor unstructured natural language requests

③ Shadow AI Tool Intelligence and Inventory

- Enhance detection capabilities to identify emerging tools such as Perplexity and DeepSeek, in addition to unofficial instances of ChatGPT and Gemini
- Establish blacklist mechanisms and detection baselines using DNS, IP, and User-Agent data—comparable to Advanced Persistent Threat (APT) prevention measures

3.2 Control Strategy

Following detection, the structural management of Shadow AI usage requires the implementation of stringent access control policies and the provision of authorized internal alternative models.

① Restrict AI Usage Permissions Based on RBAC

- Implement differentiated AI access permissions for each department using Role-Based Access Control (RBAC)
- Block external AI usage for functions such as design and R&D, while allowing only secure summarization capabilities for roles like marketing
- 'Establish and automate policies in accordance with the principle of least privilege

② Proxy-Based Blocking and AI SaaS Blacklisting

- Block access to AI services offered in a Software as a Service (SaaS) model at the HTTPS proxy layer
- Expand visibility and control by implementing automated discovery of newly emerging AI services

③ Internal Operation of Private LLM Environments

- Promote the adoption of internal AI models, such as Azure OpenAI Private Endpoint
- Preemptively control external access to maintain security governance within the organizational infrastructure boundary

④ Real-Time Sensitive Information Filtering via AI-aware DLP

- Detect sensitive data types, including PII (Personally Identifiable Information), IP (Intellectual Property), and mCAD (manufacturing CAD data)
- Leverage AI-specific DLP solutions and related products

3.3 Mitigation Strategies

The mitigation phase encompasses Zero Trust-based data flow controls, enhanced user awareness mechanisms, and the establishment of robust incident response protocols.

① Zero Trust-Based Prompt Pathway Control

- Regulate external LLM request channels through Zero Trust Network Access (ZTNA) authentication and authorization mechanisms'
- Analyze and restrict "data transmission" at each stage, from internal networks to internet gateways and ultimately to external AI services

② Security Nudging: Policy-Driven Alerts and Awareness

- Utilize KPIs (Key Performance Indicators) such as departmental AI usage frequency, detection counts, and policy violation trends
- Reinforce organizational awareness and shared accountability through regular reporting

③ KPI-Driven Monitoring and Executive Reporting

- Apply Role-Based Access Control (RBAC) for differentiated departmental permissions
- Prohibit external AI usage for design/R&D roles, while permitting only secure summarization for marketing and similar functions
- Establish and automate least-privilege policies

④ Incident Response Preparedness – Prompt Logging, Backup, and Analysis

- Integrate prompt/response log analysis within the Security Operations Center (SOC)
- Include the capability to track the scope of exposure, API usage records, and user identities in the event of an incident

4. Governance and Policy-Driven Organizational Response Framework

Technical countermeasures alone are insufficient to address the threats posed by Shadow AI. Because these risks are compounded by employee unawareness, habitual usage patterns, and the absence of robust policies, it is imperative to establish an organization-wide management framework through comprehensive security governance and informed decision-making processes.

4.1 Shadow AI Policy Framework

① AI Usage Policy

- Clearly document the criteria for prohibiting or permitting Shadow AI usage to ensure organization-wide understanding.
- The policy must specify: which AI tools may be used (maintaining allow/conditional/deny lists); what types of data are prohibited from input (e.g., PII, CAD files, design documents, source code, with illustrative examples); the disciplinary measures for violations; and procedures for exception approvals.

② AI Risk Classification (Business-Driven Risk Grading)

- Assign and manage risk levels for AI usage based on department, role, and business process.
- Establish differentiated approval and control mechanisms for each risk tier (e.g., RBAC + AI Usage Scope Matrix).

③ AI Usage Approval Process

- Require prior review by the security team or AI governance committee for any requests to use new AI tools.
- Implement technical evaluation processes for API communications, browser extensions, and internal network access requests.
- Mandate administrative approval procedures for exceptional use cases.

4.2 Education and Organizational Awareness Enhancement Strategy

More than 80% of Shadow AI incidents result from unintentional use without security awareness. Accordingly, comprehensive awareness programs—rather than simple restrictions—are indispensable at the enterprise level.

① AI Security Awareness Training Program

- Deliver regular training (at least semi-annually) and develop dissemination materials.

② Distribution of Prompt Authoring Guidelines

- Publish practical, field-oriented guides highlighting examples of “strictly prohibited prompts.

③ Operation of a Security Accountability System

- Appoint security leaders within each department to monitor AI usage, conduct campaigns, and report issues.
- Facilitate channels of communication between departments and the security team.
- Maintain continuous operation of internal security issue-sharing platforms.

4.3 AI Governance Organizational Model

① AI Risk Control Taskforce

- Composition: Security team (CISO), IT (CIO), Legal, Internal Controls, and representatives from each business unit
- Role: Manage an internal AI tool whitelist, share weekly Shadow AI detection reports, and coordinate new policies and violation responses

② AI Risk Steering Committee

- Operates as an executive reporting structure, enabling rapid decision-making in response to elevated risk levels
- KPIs: Shadow AI detection rate, number of violations, security guideline training completion rates, etc.

③ Integration with Audit and Internal Control

- Incorporate internal audit items relating to AI usage
- Regularly report on security logs, prompt usage history, and external access records

4.4 Industry Standards and Compliance Alignment

In addition to strengthening security governance within the manufacturing sector, alignment with both domestic and international legal and industry standards is essential.

Regulatory Standard	Application Area	Response Strategy
ISO/IEC 42001	Establishment of a governance framework for generative AI operations	Classification of AI risk levels; operation of oversight committees
NIST AI RMF	AI risk management framework	Inclusion of Shadow AI risk response measures
KISA AI Security Guidelines	Domestic industry-based AI security recommendations	Incorporation of AI prompt filtering and sensitive data detection
GDPR/Personal Information Protection Act	Automation processing and sensitive data leakage	Implementation of pre-input AI detection and data masking mechanisms

5. Conclusion and Response Roadmap Proposal

5.1 Conclusion

Shadow AI has rapidly emerged as a novel security risk that transcends conventional IT controls, posing direct threats to organizational confidentiality and competitiveness. This risk is particularly acute for manufacturing enterprises, where industrial trade secrets—such as design blueprints, proprietary process know-how, and cost data—are increasingly susceptible to external leakage via LLM (Large Language Model)-based AI tools.

This Insight has provided an integrated response strategy to Shadow AI threats, spanning technical, policy, and governance dimensions. The key elements of this response are as follows:

- Detection: Securing AI usage visibility through AI-aware DLP, CASB, DSPM, and related tools
- Control: Establishing AI usage policies, enforcing proxy-based blocking, and implementing role-based access control (RBAC)
- Mitigation: Controlling data pathways via Zero Trust principles, deploying alert UIs, and establishing robust incident response systems
- Governance: Instituting enterprise-wide policies, departmental risk classification, continuous education, and structured internal audits

Such measures should not be viewed as one-off policies, but rather must be embedded into organizational culture and security governance frameworks.

5.2 Proposed Response Roadmap

Outlined below is a three-phase roadmap for responding to Shadow AI:

[Phase 1: Visibility and Awareness Enhancement]

- Objective: Identify and understand the presence and risks of Shadow AI
- Key Actions:
  → Identify the existence and risks of Shadow AI
  → Conduct an internal assessment of Shadow AI usage
  → Distribute educational materials on Shadow AI incident cases
  → Establish departmental frameworks for sensitive data classification

[Phase 2: Policy and Technical Control Establishment]

- Objective: Control and minimize the use of Shadow AI
- Key Actions
  → Establish and disseminate AI usage policies
  → Configure RBAC-based AI access permissions
  → Apply and test DLP policies for sensitive data
  → Implement proxy-based blocking mechanisms for LLM access

[Phase 3: Organizational Embedding and Governance]

- Objective: Institutionalize the response framework within the organization
- Key Actions
  → Operate an AI governance committee and implement a security accountability system
  → Monitor AI usage and produce regular reports
  → Conduct ongoing AI security awareness training
  → Refine compliance response systems for AI, aligning with standards such as ISO and NIST

5.3 Future Tasks and Recommendations

- Consideration of Internal LLM Deployment: Establish private LLM environments to leverage generative AI capabilities without incurring security risks, thereby reducing reliance on external Shadow AI services.
- Expansion of AI-Specialized Security Solutions: As existing security appliances struggle to detect the unstructured nature of LLM interactions, it is essential to adopt AI-aware DLP, prompt security filtering, and data flow detection technologies.
- Evolution of the Security Team's Role: Responding to Shadow AI threats requires security teams to transition from mere monitoring to serving as AI utilization advisors and security consultants.
- Advancement of Legal and Regulatory Compliance Systems: With generative AI-related regulations evolving rapidly, dedicated organizational structures and the integration of audit criteria are necessary to ensure compliance.

Shadow AI is not merely a matter of technological adoption, but a security imperative that fundamentally determines the protection of trade secrets and, ultimately, the survival of the organization. It is now essential to implement multilayered countermeasures—spanning technology, policy, and culture—in an integrated manner.

If your organization requires the development of security policies to safeguard industrial trade secrets from Shadow AI threats, we encourage you to leverage SK Shieldus's extensive expertise in technology and policy to initiate a robust AI security governance framework.

■ References

[1] Structured, Shadow AI – The Hidden Threat to Governance & Compliance, 25.04

[2] Inteleca, Shadow AI in the Workplace: The Hidden Security and Compliance Risks, 25.03

[3] CIODIVE, AI-generated code leads to security issues for most businesses, 24.01

[4] Nightfall AI, The Nightfall Approach: 5 Ways Our Shadow AI Coverage Differs from Generic DLP, 25.07

[5] NIST AI RISK MANAGEMENT FRAMERK (AI RMF), 23.01

■ Additional Resources

[1] Paloalto, What Is Shadow AI? How It Happens and What to Do About It (Cyberpedia)

[2] ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system

[3] Ministry of the Interior and Safety(South korea), AI Security Guidelines for Public Institutions, Oct. 2023

[4] NIPA, Report on Generative AI Utilization and Security Threats by Industry, 2024

[5] SK Shieldus, EQST Insight Blog Series (2023–2024)