“Shadow AI” - Your next data breach might already be around the corner

Andreas Mayer, Senior Principal Architect for Northern Europe, Google Cloud

Thiébaut Meyer, Office of the CISO, Google Cloud

Google Cloud

Table of contents

Executive Summary

Generative Artificial Intelligence (GenAI) is no longer a future technology, but an integral part of your employees' daily workflow. This adoption is happening with or without official sanction, creating a phenomenon widely known as "Shadow AI," where employees use consumer-grade tools for business purposes, often undermining significant investments in cybersecurity. This trend introduces severe risks, including data leakage, privacy violations, and non-compliance. We wrote about these specific aspects in the Google Cloud Blog post Spotlighting 'shadow AI': How to protect against risky AI practices.

However, the challenge has evolved beyond just consumer tools. A newer, more subtle form of Shadow AI has emerged: the use of enterprise-grade AI platforms without proper corporate oversight. This "Business AI Gone Rogue" is often driven by a desire for innovation and efficiency, but it operates outside of established governance, creating security gaps and a chaotic "AI Tool Zoo". More details can be found in the Google Cloud Security Community post Shadow AI Strikes Back: Enterprise AI Absent Oversight in the Age of Gen AI.

The proliferation of unsanctioned AI models and tools already presents a significant challenge for enterprise governance and security. However, this landscape is rapidly evolving. The risk is no longer confined to unauthorized data analysis but is escalating to include what can be called *Shadow Agents*. As detailed in the Google Cloud Security article, Shadow Agents: A New Era of Shadow AI Risk in the Enterprise, these are autonomous or semi-autonomous systems, often built by employees, that can execute tasks, access data, and interact with other systems entirely without IT oversight. This evolution from unsanctioned tools to unsupervised agents introduces a more dynamic and unpredictable threat vector, magnifying the potential for data exfiltration, compliance breaches, and operational incidents.

Building on this context, this whitepaper focuses on the most critical and immediate consequence of all forms of Shadow AI. The core problem is the uncontrollable leakage of sensitive data. With every prompt, your employees risk transferring intellectual property (IP), personally identifiable information (PII), and other critical corporate data into external, uncontrolled environments. As these models may be trained on user data, the future use and whereabouts of your data are uncertain. This not only constitutes an acute security risk but also creates significant compliance violations, in particular under regulations like GDPR and the EU AI Act.

This document analyzes the profound impact of this phenomenon on your organization and demonstrates that ignoring this trend is not an option. Navigating this landscape requires a nuanced strategic approach, especially for organizations operating in regulated environments where both preventative measures and, when necessary, proportionate disciplinary actions are critical for compliance.

You are therefore facing a strategic, cumulative choice with massive consequences:

- **Exclusive Prohibition and Repression**: While appearing straightforward, an exclusive reliance on banning tools and punitive measures alone often proves short-sighted. This approach can stifle innovation, erode trust, and inadvertently push shadow usage deeper into unmanageable territory.

- **Proactive Integration and Governed Enablement**: This strategy focuses on securely leveraging GenAI's immense productivity. It involves integrating AI tools responsibly within your organizational framework through robust governance, regaining control, and ensuring continuous legal compliance.

- **Integrated Risk Management**: The most robust and recommended approach, particularly for regulated enterprises. This strategy synthesizes proactive integration and comprehensive governance with essential preventive controls and, where appropriate, clearly defined and proportionate repressive actions. This ensures a balanced framework that fosters innovation while rigorously meeting compliance obligations.

After reading this document, you will be equipped with the foundation you need to understand these threats and make an informed, strategic decision. This is not an IT problem but a strategic business imperative.

The Threat: Shadow AI as an uncontrolled data gateway

GenAI tools offer a simple yet powerful solution for summarizing texts, generating code, drafting emails, and developing concepts. The potential is massive, the appeal immense, and access effortless. From a corporate security perspective, as this can introduce uncontrolled data exfiltration events, it is **a significant problem**[1][2].

Your existing security measures are often incapable of analyzing the contextual content of data sent to APIs. They may see encrypted HTTPS traffic to a legitimate service but cannot discern that the content is a draft for a patent application or a list of customer data.

Even if the access to such services is directly blocked from your Enterprise IT network, your employees still have private devices such as smartphones, tablets, laptops, etc. that they can use to generate results for them as intermediaries. The process of data input and leveraging of the results then requires a higher degree of manual effort, but it might still be worth it from the perspective of an employee.

Risk Scenarios: Direct Data Leaks

To make the danger more tangible, we will consider three types of unintentional, employee-driven data leaks.

Scenario 1: The accidental GDPR violation

An HR manager wants to quickly summarize a candidate's cover letter. They copy the entire text, including the name, address, and contact details into a consumer grade AI tool. At that moment, Personally Identifiable Information (PII) is transferred to a third party provider, potentially outside the EU, without a legal basis, without a Data Processing Agreement (DPA), and without the data subject's knowledge. This is a clear violation of the GDPR, which can lead to substantial fines. It would however be very hard to discover such a violation.

Scenario 2: The Theft of Intellectual Property

Your engineer is working on an innovative technical solution. To optimize a piece of source code, she pastes it into a consumer grade GenAI tool and prompts the system to receive suggestions for improvements. This code, representing the company's core intellectual property (IP) and the heart of a future product, now resides on an external provider's servers and could potentially be used to train their global models. Your competitive advantage is now at risk[3][4].

---

[1] https://www.cybsafe.com/press-releases/study-almost-40-of-workers-share-sensitive-information-with-ai-tools-without-employers-knowledge/
[2] https://cloudsecurityalliance.org/blog/2025/03/04/ai-gone-wild-why-shadow-ai-is-your-it-team-s-worst-nightmare#
[3] https://www.ciodive.com/news/Samsung-Electronics-ChatGPT-leak-data-privacy/647137/
[4] https://www.entrepreneur.com/business-news/apple-bans-employee-chatgpt-use-over-data-privacy-concerns/452520

The Next-Level Threat: Data Poisoning and Sabotage

Beyond unintentional leaks hides a more sophisticated and malicious threat: the deliberate manipulation of AI models, a practice known as **Data Poisoning** or **Adversarial Machine Learning**. This risk is particularly relevant when considering nation-state actors or ruthless competitors.

While these threats are independent from the question whether a solution or tool is used in a legitimate way or via Shadow AI, there is still a significant difference with regards to the potential risks involved. When enterprises select the software and services that they plan to deploy as part of their Enterprise IT, they usually conduct a proper Vendor Risk Management (VRM) and perform security audits on the specific products and technologies to understand which risks exist, how they can be mitigated, and which risks need to be accepted.

In order to create a better understanding of this complex threat, we again leverage three easy to understand scenarios.

Scenario 1: Cybersecurity and Critical Infrastructure Sabotage

A nation-state actor poisons the training data of a popular code-assistance AI. The model is trained to introduce a subtle but critical vulnerability, a backdoor, when asked to generate code for specific functions like encryption or network protocols. Developers in hundreds of companies and government agencies in a target country adopt this code in good faith, unknowingly creating a widespread, exploitable weakness for espionage or disruption.

The distinction between Shadow AI and a sanctioned deployment is fundamentally different here:

1. **Secure Supply Chain Integration:** In a governed enterprise environment, any code-assistance AI is treated as a component of the software supply chain. It would be vetted, and its outputs would be subject to the same rigorous automated scanning, security analysis, and peer review as human-written code. Shadow AI completely bypasses this Secure Software Development Lifecycle (SSDLC).

2. **Verifiable Integrity**: A sanctioned enterprise solution would ideally come with a verifiable "AI Bill of Materials" (AIBOM) or model attestation, providing assurance about its training data and integrity[5]. With Shadow AI, developers are using a black box with no provenance, making it impossible to trust the output.

3. **Centralized Incident Response**: If a backdoor in the sanctioned AI were discovered, a central security team could immediately disable the tool, use logs to identify every single instance of code it generated, and deploy a targeted remediation. In the Shadow AI

[5] https://cloud.google.com/blog/topics/threat-intelligence/securing-ai-pipeline/?e=48754805

scenario, the organization is blind and it has no record of which developers used the tool or where the vulnerable code now resides, making a full cleanup nearly impossible.

With Shadow AI, you are not just using a potentially flawed tool, but you are allowing your entire software supply chain to be compromised by an invisible, untraceable, and ungoverned threat vector.

Scenario 2: Economic and Corporate Sabotage via Ungoverned AI

Your competitor intentionally seeds public forums and code repositories with slightly flawed but plausible technical data related to your R&D focus, a classic data poisoning strategy[6][7].

When your engineers use unvetted, consumer-grade Shadow AI tools for research, the model confidently presents this poisoned data as fact. Because these public models are not grounded in your organization's proprietary, trusted data, and have not been vetted through a Vendor Risk Management (VRM) process, your team has no mechanism to validate the information's integrity. They accept it at face value. This leads your R&D team down the wrong path, causing project delays and significant financial losses.

The distinction between Shadow AI and a sanctioned deployment is critical here. In a governed enterprise environment, this risk would be mitigated in several ways:

1. **Model Vetting**: A sanctioned AI platform would have undergone a rigorous security review to assess its defenses against data poisoning.

2. **Data Grounding**: The enterprise AI would be configured to prioritize or be exclusively grounded in your own internal, trusted knowledge bases and R&D data. When faced with conflicting information from the public domain, it would flag the discrepancy or rely on the verified internal source.

3. **Centralized Oversight**: If a sanctioned model began providing flawed outputs, a centralized AI governance team could immediately investigate, disable the model for the entire organization, and prevent further damage.

With Shadow AI, you have none of these safeguards. Each employee is using a black box, and the organization is blind to the poisoned data corrupting its strategic decision-making. The failure isn't just that the AI was wrong, but that the lack of governance created an environment where the lie was undetectable.

Scenario 3: Geopolitical Influence and Corporate Decision-Making

An actor seeking to shape public opinion systematically poisons the knowledge base of major, public-facing LLMs with propaganda or biased versions of historical or political events.

[6] https://arxiv.org/abs/2302.10149
[7] https://www.wiz.io/academy/data-poisoning

Within your organization, a market analyst or strategy team uses a popular but unvetted Shadow AI tool to quickly generate a report on the sociopolitical landscape of a potential new market. The AI confidently outputs the manipulated narrative, presenting it as objective fact. Your team, trusting the tool's apparent authority and lacking any way to audit its sources, unknowingly bases a critical market expansion, partnership, or investment decision on state-sponsored disinformation.

The distinction between this and using a sanctioned enterprise AI is significant:

1. **Source Transparency and Grounding**: A governed enterprise AI would be configured to draw from and cite trusted, vetted sources (e.g., specific news agencies, economic journals, proprietary geopolitical risk reports). It would provide auditable sources for its claims, allowing the team to verify the information. Shadow AI offers a confident answer with no verifiable "receipt," forcing blind trust.

2. **Customizable Guardrails**: An enterprise platform would have guardrails tailored to your organization's risk profile. It could be configured to flag or refuse to answer queries on sensitive geopolitical topics, instead directing the user to a human expert or a curated, trusted data source. Public tools have generic guardrails, but they are not aligned with your specific corporate governance needs.

3. **Monitoring and Central Control**: If it's discovered that a sanctioned AI provided a biased analysis, a central team can instantly check the system, identify every employee who received that information, and issue a correction. In the Shadow AI scenario, the bad information spreads silently throughout the organization with no audit trail, poisoning strategic conversations and decisions without anyone knowing the source.

With Shadow AI, you are not merely risking an inaccurate report but are effectively allowing anonymous, and potentially malicious, external actors to influence your most critical business strategies. You are outsourcing your corporate intelligence to an untrusted, ungoverned, and manipulative source.

Further Operational and Strategic Risks

Beyond the technical risks, the very way your company operates and competes is put at risk. To ensure a broad applicability of the examples, we want to share a few additional impact areas.

Enterprise Architecture: The carefully designed systems for data governance, integration, and sovereignty are rendered irrelevant. Data flows become chaotic and untraceable, creating a shadow system that operates outside of all established and proven architectural principles. In the absolute worst case scenario you can think of this as a collapse of parts of your enterprise architecture.

Erosion of Process Integrity: Business processes, from product development to financial reporting, rely on being repeatable, auditable and reliable. When employees use unverifiable AI tools to complete tasks, processes become non-standard and impossible to audit. A critical calculation or design step performed by an external AI cannot be validated or reproduced, which introduces dangerous elements of unpredictability into your operations.

The Regulatory Challenge

The unmanaged use of Shadow AI does not exist in a vacuum. It represents a direct collision with a dense and rapidly maturing landscape of laws, regulations, and industry standards. In several jurisdictions, a legal framework on AI has completed the existing regulations on data. It is in particular the case in the European Union with the publication of the EU AI Act (Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024) in July 2024. Allowing Shadow AI to persist is not just a security risk, but a profound compliance gap that may expose the organization to severe legal, financial and reputational damage.

GDPR (General Data Protection Regulation)

The GDPR is the foundation of data protection in Europe. The use of a shadow AI system within an organization may pose significant compliance challenges with this regulation.

For instance, the GDPR lists in its Article 5 several core principles relating to processing of personal data, the first being lawfulness, fairness, and transparency. Let’s consider a marketing department that uses a shadow AI system to analyze customer browsing history and create targeted advertising. Customers are not informed about this data processing. This poses challenges with the transparency and fairness principles. A shadow AI system processing personal data without the data subject's knowledge or consent, and outside of the organization's established data processing policies, may violate this first principle. And similar risks exist with the other principles.

Another example of compliance risk under the GDPR with the use of shadow AI would be the requirement that processed data are protected by design and by default (Article 25). Implementing data protection by design and by default becomes practically impossible with shadow AI, as the privacy frameworks of the consumer-oriented services often used are inherently misaligned with business requirements. The security assurances, data minimization capabilities, and privacy controls necessary for a corporate environment are simply not the focus of systems built for individual use..

Finally - and the list could easily be extended - shadow AI makes it difficult to run a proper Data Protection Impact Assessment (DPIA). The GDPR mandates such assessment under Article 35 when data processing activities are likely to result in a high risk to the rights and freedoms of individuals. This assessment is a critical compliance tool, designed to proactively identify and mitigate potential privacy risks before data processing begins. However, the very nature of shadow AI—its clandestine and unauthorized deployment—renders a comprehensive DPIA impossible.

EU AI Act

The use of a shadow AI system creates a substantial risk of meeting the requirements of the EU AI Act, as the Act's entire framework is built on a foundation of risk assessment, transparency, and documented compliance - principles that are fundamentally incompatible with unauthorized, unvetted systems.

For example, consider a human resources department using a shadow AI tool to automatically screen and rank job applicants. This system would be considered as 'high-risk' under Article 6 (in conjunction with Annex III point 4). By its very nature, a shadow AI system makes compliance with high-risk requirements impossible, implicating both the system itself and its improper usage by an organization. The system inherently lacks the provider-mandated risk management framework (Article 9), data governance (Article 10), and technical documentation (Article 11). This deficiency, in turn, prevents the deploying organization from fulfilling its own crucial obligation to apply appropriate human oversight to prevent or minimize risks (Article 14).

Similarly, a legal tech team in a law firm wants to create a tool to help with case preparation. They use a powerful, publicly available (but unvetted by the firm) General-Purpose AI (GPAI) model via a shadow subscription. They build a system on top of it that summarizes legal documents and identifies precedents. They are unaware of the GPAI model's training data, its known limitations, or its potential for hallucinating non-existent case law. This may create a critical downstream compliance failure. While the primary obligation to provide documentation for the GPAI model lies with its original provider, the law firm, by using this model to build their own system, has a responsibility to understand its components.

In both scenarios, the lack of formal oversight and risk management not only breaches these specific articles but also exposes the company to severe penalties, making the use of shadow AI a critical compliance failure under the new regulation.

EU Data Act

The EU Data Act establishes a legal framework giving users (both individuals and companies) of connected products the right to access the data they generate and to share it with third parties of their choice. It clarifies data ownership and access.

When an employee uploads proprietary IoT data or user-generated data to an external AI service for analysis, they create a new, uncontrolled data silo. This fundamentally complicates your ability to comply with data access requests from your own customers, as required by the Act. You no longer have a clear chain of custody for the data you are legally obligated to manage.

In conclusion, from a legal and compliance perspective, Shadow AI places the organization in a state of gross negligence. It is not a legal gray area, it is a real challenge to comply with binding

laws and principles of established compliance practices. Claiming back control is highly important.

How Google Cloud supports a Governed Approach to tackle Shadow AI

Transforming the risks of Shadow AI into a controlled, strategic asset requires a holistic approach that integrates policy, people, and platform. Rather than building a program from scratch, organizations can leverage established, world-class frameworks and best practices to accelerate their journey to secure AI adoption.

The cornerstone of this strategy is a robust governance framework. Google's Secure AI Framework (SAIF) provides a comprehensive blueprint for securing the entire AI lifecycle, from data collection to model deployment. The whitepaper SAIF in the Real World focuses on “Key considerations in applying the Secure AI Framework (SAIF) through the AI development lifecycle”. This framework helps organizations implement practical, essential steps such as establishing a clear AI Acceptable Use Policy and adopting systematic risk management, guided by globally recognized standards like the NIST AI Risk Management Framework.

Securing the AI Software Supply Chain is also a topic of the highest relevance in the context of developing AI systems through their whole life-cycle and this very insightful whitepaper provides an excellent starting point.

However, a framework is only effective if people follow it and if it is underpinned by a sound, company-overarching compliance culture. An appropriate compliance culture ensures the framework is more than just a document, it becomes a living part of the organizational mindset. It encourages a proactive approach where employees understand that data security is a collective duty, not just the IT or legal department's problem. In such an environment, policies are not seen as restrictive hurdles, but as enabling guardrails that guide innovation safely. Proactive and continuous employee training is critical. This means educating teams on both the why (the risks of PII and IP leakage) and the how, including safe and effective prompting.

Finally, the chosen technology platform must enforce policy and empower users securely. Google Cloud's AI solutions, like the Vertex AI Platform, are designed with these principles at their core, offering enterprise-grade access controls, data governance features, and integrated tools for the continuous monitoring and auditing essential for a mature AI program.

Architecting for the AI era

The challenges described in this paper like data leakage, model poisoning, and loss of competitive advantage, cannot be solved by simply banning tools or through employee training alone. The strategic imperative is to provide an enterprise-grade AI platform that makes the secure path the easiest path.

However, in an era of rapid innovation, organizations rightfully fear being locked into a single technology. The ideal platform must therefore be built on a foundation of core architectural principles that ensure security, prevent lock-in, and empower users.

From our perspective, an enterprise-grade AI platform that is fit for the Agentic Era must be built on three foundational pillars:

1. Centralized Governance and Security by Design

This is the direct antidote to Shadow AI. A governed platform provides a single pane of glass for security, risk, and operations. It enforces the AI Acceptable Use Policy by design, with granular Identity and Access Management (IAM), data encryption, and robust audit logs. It allows organizations to manage AI as a core business function, not an uncontrollable risk.

- *This principle is embodied in platforms like Google's Vertex AI, which centralizes model management, access controls, and monitoring to provide a secure foundation for all AI development.*

2. Openness and Model Choice

To avoid vendor lock-in and harness the best innovation, a platform must be fundamentally open. It should provide access to a diverse ecosystem of models, including state-of-the-art proprietary models, open-source alternatives, and the organization's own custom-built models. This ensures that the enterprise is always using the best tool for the job, without being tied to a single vendor's roadmap.

- *The Vertex AI Model Garden, for example, is built on this philosophy, offering access to Google's models alongside dozens of third-party and open-source options in a single, managed environment.*

3. Pervasive Integration Where Work Happens

Shadow AI thrives when official tools are siloed or hard to use. A successful platform must embed AI directly into the daily workflows of every employee. It should be a natural extension of the productivity suites, development environments, and business applications they already use. When AI assistance is contextual and readily available, the incentive to seek out ungoverned, external tools disappears.

- *This deep integration is the core idea behind Gemini in Google Workspace, which brings generative AI capabilities directly into documents, email, and spreadsheets, meeting users exactly where they are.*

By selecting and building upon a platform founded on these principles, an organization can resolve the central paradox of the *Agentic Era*. It can empower its teams with cutting-edge AI to drive innovation and competitive advantage, while simultaneously transforming the existential risk of Shadow AI into a governed, resilient, and strategic asset.

To facilitate the rapid innovation of AI solutions, we published a blog post illustrating how Google Cloud can specifically assist users throughout their journey: Introducing AI Protection: Security for the AI era.

Conclusion: Navigating from Dilemma to Strategy

The drive for competitive advantage is precisely what makes the topic of AI so critical and, paradoxically, what fuels the proliferation of Shadow AI. Employees, in their pursuit of efficiency and innovation, are naturally drawn to powerful AI tools. Organizations that fail to embrace and integrate AI will undoubtedly face competitive stagnation, losing ground to more agile competitors.

However, the risk lies not in the adoption of AI itself, but in its uncontrolled deployment. When proprietary algorithms, R&D data, or go-to-market strategies are fed into unvetted, consumer-grade AI tools, you are effectively gifting your competitive advantage to an unknown third party. This leakage doesn't just represent a single data breach but represents the irreversible erosion of your unique market position.

This situation presents leadership with a complex strategic challenge, rather than a simple binary choice. It is not merely a question of banning AI (and thus falling behind) versus allowing unmanaged Shadow AI (and risking intellectual property). The reality is that the inherent human desire for efficiency means that Shadow AI risks can persist even when sanctioned enterprise solutions are provided, necessitating a broader approach.

The true strategic imperative is therefore to move beyond this false dilemma by embracing comprehensive governance. This proactive approach involves not only providing secure, powerful, and compliant GenAI solutions but also proactively managing employee behavior through robust training, clear policies, and continuous enablement. By transforming unmanaged AI use into a controlled asset, you convert a source of competitive risk into a sustainable advantage, ensuring innovation while safeguarding your most valuable data.

The only viable and strategically sound path is **proactive and preventative governance**. By providing your employees with a secure, powerful, and compliant GenAI solution, you channel their drive for innovation into productive and controlled avenues. You transform an incalculable risk into a calculated competitive advantage.

We suggest the following three next steps:

1. **Discover and Assess Your Current AI Footprint**. You cannot manage a risk you cannot see. Initiate a rapid assessment to understand your organization's current Shadow AI usage. Combine network traffic analysis to identify popular external AI services with anonymous surveys to gauge employee needs and behaviors. This data will provide a clear, evidence-based picture of your actual risk exposure and the business problems your employees are trying to solve.

2. **Establish an AI Governance Baseline**. Assemble a cross-functional AI council with leaders from Security, Legal, HR, and key business units. Your first objective is to draft a version 1.0 of your "AI Acceptable Use Policy." This document should not aim for

perfection but should establish clear, simple guardrails on data classification (e.g., "Public," "Internal," "Confidential") and define which types of data are strictly forbidden from being used in any external, unvetted tool.

3. **Launch a Controlled, High-Value AI Pilot**. Instead of attempting to boil the ocean, select a single, high-impact business process (e.g., summarizing internal research, drafting marketing copy, or analyzing non-sensitive operational data) for a governed AI pilot. Use this controlled environment to evaluate a secure, enterprise-grade AI platform against the governance policy you've established. This allows you to demonstrate tangible value, refine your security controls, and build a strong business case for a broader, secure rollout.

The decision is yours: Will you leave control over your most valuable data to chance, or will you actively shape it? Google Cloud is your reliable partner on this exciting journey. The time to act is now - Empower Your AI Adoption with Google's Approach to Responsible Innovation.