ORION

CISO TRADECRAFT

WHITE PAPER

POLICY-RELIANT DLP FAILS AGAINST SHADOW AI

CO WRITTEN BY
G. MARK HARDY (CISO TRADECRAFT) & DANIEL BASS (ORION)

POLICY-RELIANT DLP FAILS AGAINST SHADOW AI

EXECUTIVE SUMMARY

Data Loss Prevention (DLP) based on policies cannot keep pace with Artificial Intelligence (AI).

We write policies that appear sound on paper and translate them into DLP rule sets, yet we constantly struggle with false positives and missed incidents. The former blocks legitimate business and places cybersecurity at odds with the organization; the latter represents a failure to meet the basic requirements of DLP - neither option is acceptable in today's environment of regulation, the increased threat of data loss due to AI-automation, and user misunderstanding of the consequences of sharing sensitive information with AI.

The rapid increase in the use of generative AI tools is evident. It's driven by employee adoption, the new native capabilities of SaaS platforms, or, in some cases (like Microsoft Copilot) enabled by default, and is rarely preceded by a security review or governance.

This phenomenon, commonly referred to as "Shadow AI," poses a significant yet poorly understood data-loss risk for organizations.

This paper examines why the traditional, policy-based approach to DLP fails in the age of AI.

It examines how data is leaking into unauthorized AI systems, why traffic monitoring alone is insufficient, and how security leaders can protect sensitive information without impeding business operations. Finally, it proposes a modern, context-aware approach designed to meet the realities of AI-enabled workflows.

IN THE BEGINNING…

The original DLP strategy was a "dirty word list" — a simple static list matched to specific values of content flowing past a chokepoint. The US military maintained networks at various classification levels (unclassified, secret, top secret), but information occasionally needed to flow from one level to another. By filtering content based on known keywords (e.g., "Project Secret"), inadvertent disclosures could be prevented. A simple character string match was often sufficient for this task. Beyond text, the hash values of known sensitive files could be matched, triggering an exception to the transfer of attachments. This worked great when all content was unencrypted and moving at 1990's speed, but its lack of flexibility meant that constant updates, usually time-late, created a game of catch-up that often missed new content (e.g., Project New-Stuff) while triggering false positives (e.g., "we held a confidential meeting with the client") These early efforts were important, but failed to meet developing requirements.

THE AWAKENING: CONTENT-AWARE INSPECTION

Privacy regulations such as HIPAA, GLBA, and SOX began to be enforced in the early 2000s, increasing the consequences of DLP failures. Solutions evolved with requirements, using tools such as regular expression matching (regex) to validate credit card numbers, social security numbers, and other sensitive fields based on their format and content. This more nuanced approach enabled rules that encompassed more than static values. However, false positives were common (e.g., nine-digit Marriott guest numbers led to hotel reservations being rejected). In addition, most traffic continued to traverse networks and the internet unencrypted, making interdiction easier but also interception more likely. Furthermore, there was no universal word list for all types of information requiring protection. A massive shift in communication was about to render this approach obsolete.

POST-SNOWDEN ENCRYPTION

In 2013, Edward Snowden released a trove of classified material revealing that national intelligence agencies were intercepting vast quantities of unencrypted communications. Encryption was often optional, and confusion and complexity regarding management (along with prices charged for simple certificates) helped keep the world's secrets in the clear. As late as 2015, less than 10% of the largest one million websites offered HTTPS by default.[1] Companies (and individuals) became aware of the risk of sharing information without protection, and the gradual yet nearly ubiquitous adoption of end-to-end encryption ushered in a new era of DLP challenges.

Web gateway tools like Blue Coat became useful for their ability to "proxy" communications. Instead of an end user connecting directly to a server by exchanging HTTPS certificates, the endpoint was instructed to trust the proxy certificate instead. Now, the encrypted traffic was forwarded from the endpoint to the proxy gateway, where it was decrypted, analyzed, and reencrypted via a second end-to-end session between the gateway and the server. This was effective unless your proxy server was compromised; in that case, all of your data was compromised as well. However, these solutions did nothing to prevent an insider from inserting a USB drive into the port and directly extracting sensitive information. As early as 2004, Microsoft offered methods to disable USB storage devices, albeit through strategies that weren't particularly user-friendly, such as editing the registry.

CASB TO THE RESCUE?

As enterprises migrated to the cloud, the problem of server-to-endpoint information transfer became a critical challenge—the "perimeter" was gone, and guarding the entry and exit points of the corporate network was ineffective. Gartner introduced the term Cloud Access Security Broker (CASB) in 2012[2] as "security policy enforcement points ... to combine and interject enterprise security policies as the cloud-based resources are accessed." Unfortunately, agent-based CASB tools were rarely installed on non-corporate devices, leaving BYOD environments unprotected. Fortunately, CASB could detect when employees were using unauthorized file-sharing tools and inspect data at rest in the cloud.

THE STATE OF DLP TODAY

The vast majority of solutions today rely on regex-based policies for DLP. Security analysts manually create these policies to try to predict and prevent potential exfiltration incidents (sources, data, and destinations) before they occur.

This idea in itself is already failing in the context of modern data and scale. The variety of data, sources, destinations, and contexts for each data trace cannot be manually translated into effective DLP policies that actually prevent data exfiltration. Moreover, the mere attempt to do so results in extremely high operation costs and an endless stream of false positives familiar to any security professional.

DLP was already failing before AI came into the picture, and its introduction made the problem even more complex.

THE AI CHALLENGE

AI presents both opportunities and challenges for securing and protecting enterprise data against loss, compromise, or inadvertent disclosure.

Microsoft 365 Copilot was automatically enabled for Windows 10 and Windows 11 licenses beginning in September 2024, with voice chat beginning in November 2025.[3] A productivity enhancer for many, it also exposed the shortcomings of prior manual data classification and access rules. SharePoint directory permissions were honored (and exploited) by Copilot to offer more complete responses - but if an intern was somehow included in a group that had access to quarterly financials, executive compensation, or internal engineering research, Copilot would merrily serve up that information when asked. Years of managing data access through simple rules that were updated infrequently, if at all, exposed a deep-rooted weakness in the enterprise security strategy. More sophisticated prompt engineering circumvents many access controls in multiple AI tools, rendering static policies ineffective at best, and offering a false sense of security at worst. However, there is an even more insidious risk that is working its way through the enterprise: Shadow AI.

---

[1] Felt, et al., Measuring HTTPS Adoption on the Web, Paragraph 6.1, accessed 24 January 2026 at https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/46197.pdf

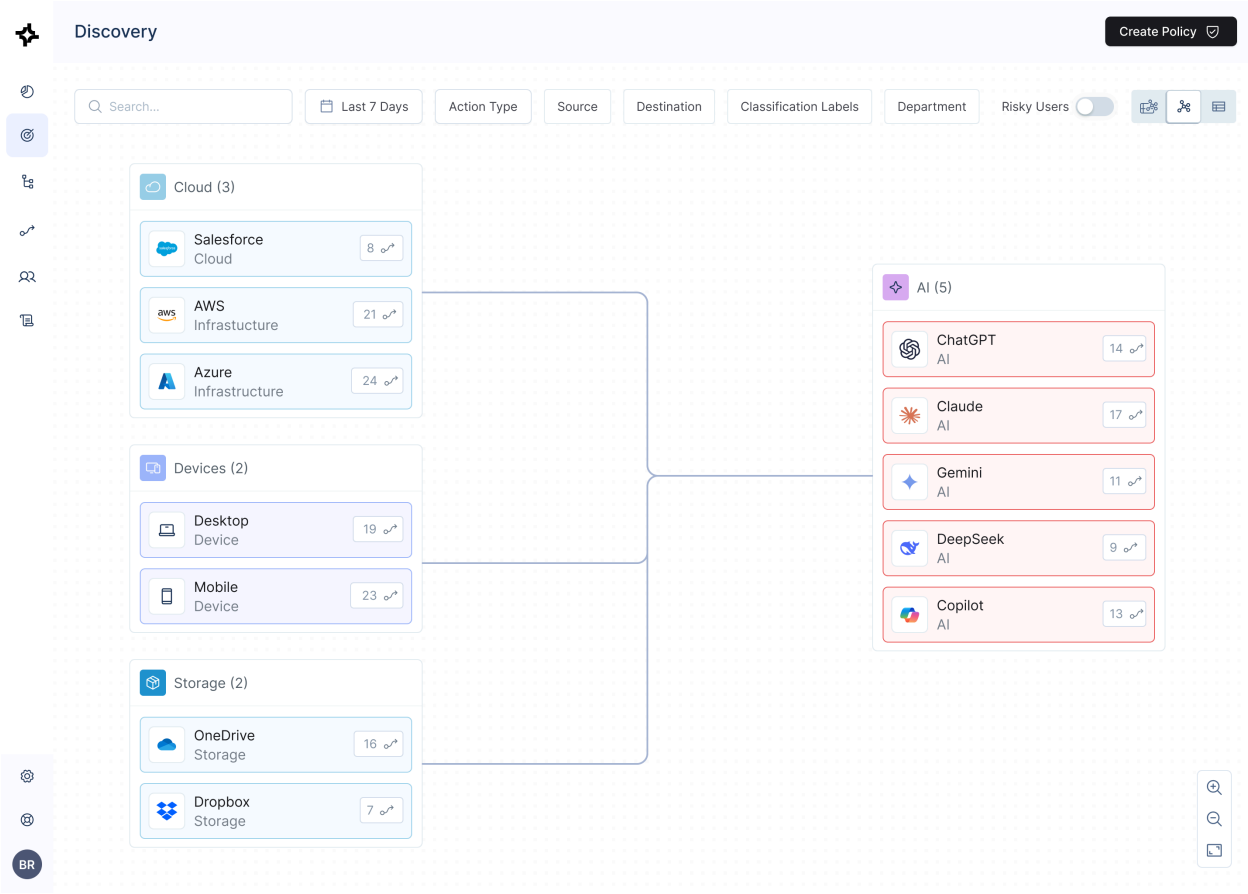
[2] "Definition of Cloud Access Security Brokers (CASBs) - Gartner Information Technology Glossary", Gartner, accessed 24 January 2026 at https://www.gartner.com/en/information-technology/glossary/cloud-access-security-brokers-casbs

[3] https://learn.microsoft.com/en-us/windows/client-management/manage-windows-copilot

THE RISE OF SHADOW AI

IBM defines **Shadow AI as "the unsanctioned use of any artificial intelligence (AI)** tool or application by employees or end users without the formal approval or oversight of the information technology (IT) department."

Shadow AI is not inherently malicious. Employees use it to summarize documents, write code, analyze data, and increase productivity. The risk arises when sensitive enterprise information is introduced into systems with which the organization has no contractual relationship regarding data ownership. Sending this information to a destination the organization does not control, cannot effectively monitor, and may not be able to audit after the fact creates a scenario in which relying on traditional DLP tools is a recipe for failure.



HOW DATA LEAKS INTO AI SYSTEMS

Unlike traditional exfiltration scenarios, AI-driven data leakage often occurs through normal business activity. Employees paste internal documents into public AI tools to generate summaries. Developers submit proprietary source code for debugging assistance. Marketing teams upload customer data to draft content. AI copilots index internal repositories and respond to prompts using information far beyond what users were intended to access.

These interactions frequently bypass existing DLP controls because the data is neither sent as an attachment nor does it match predefined patterns. It is embedded in conversational prompts, transmitted over encrypted channels, and processed by systems outside the enterprise boundary.

Once data enters an external AI system whose Terms of Service and Privacy Policies have not been formally reviewed and accepted, control is effectively lost. Organizations may have no visibility into how long the data is retained, how it is used for model training, or whether it can be retrieved or deleted.

[4] Krantz, Jonker, McGrath, "What is shadow AI?", IBM, accessed 3 February 2026 at https://www.ibm.com/think/topics/shadow-ai

THE BUSINESS REALITY: ENABLEMENT VS PROTECTION

Unlike traditional exfiltration scenarios, AI-driven data leakage often occurs through normal business activity. Employees Many organizations respond to AI risk by attempting to block AI tools outright, but this approach rarely succeeds. Employees quickly find workarounds using personal devices, home networks, or alternative platforms. Productivity suffers, and security teams lose what little visibility they had.

Conversely, allowing unrestricted use of AI exposes the organization to potential regulatory violations, intellectual property loss, and reputational damage. Data protection regulations such as the General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA) d**o not distinguish between malicious exfiltration and inadvertent disclosure.** From a compliance perspective, the outcome is the same.

Security leaders are therefore faced with a problematic choice: disrupt the business or accept unnecessary risk. Traditional monitoring strategies attempt to walk this line by observing traffic and reacting after the fact - a model which fails in the context of AI.

WHY MONITORING TRAFFIC IS NO LONGER ENOUGH

Monitoring assumes that security teams can inspect data flows, identify violations, and intervene before damage occurs. AI interactions undermine these assumptions because AI prompts are dynamic and contextual. Sensitive data may be embedded within otherwise benign conversations. Without proxying, encrypted communications limit inspection, while the speed and volume of AI-driven workflows exceed the capacity of manual review or static rule enforcement.

Even when violations are detected, they are often discovered too late. Once data has been submitted to an external AI system, blocking subsequent traffic does not undo the disclosure.

Effective protection requires the ability to understand what data is being shared, why it is being shared, and whether that action aligns with business intent, in real time.

A MODERN APPROACH TO SECURING AGAINST SHADOW AI

Addressing Shadow AI requires a shift away from static policies and toward context-aware controls.

Modern solutions address the inherent limitations of legacy DLP exposed by AI by leveraging it to analyze data flows across endpoints, SaaS platforms, email, storage, and web interactions. By learning what "normal" business behavior looks like, these systems can identify anomalous or risky activity **as it occurs** rather than after the fact.

Automatic, AI-based classification of structured and unstructured data enables decisions based on sensitivity rather than format. Behavioral analysis distinguishes legitimate use from improper exfiltration attempts, enabling targeted intervention to prevent data loss without disrupting legitimate work, while continuously learning and improving accuracy over time. Rather than relying on predefined rules that quickly become obsolete, context-aware, AI-based systems adapt continuously as workflows evolve.

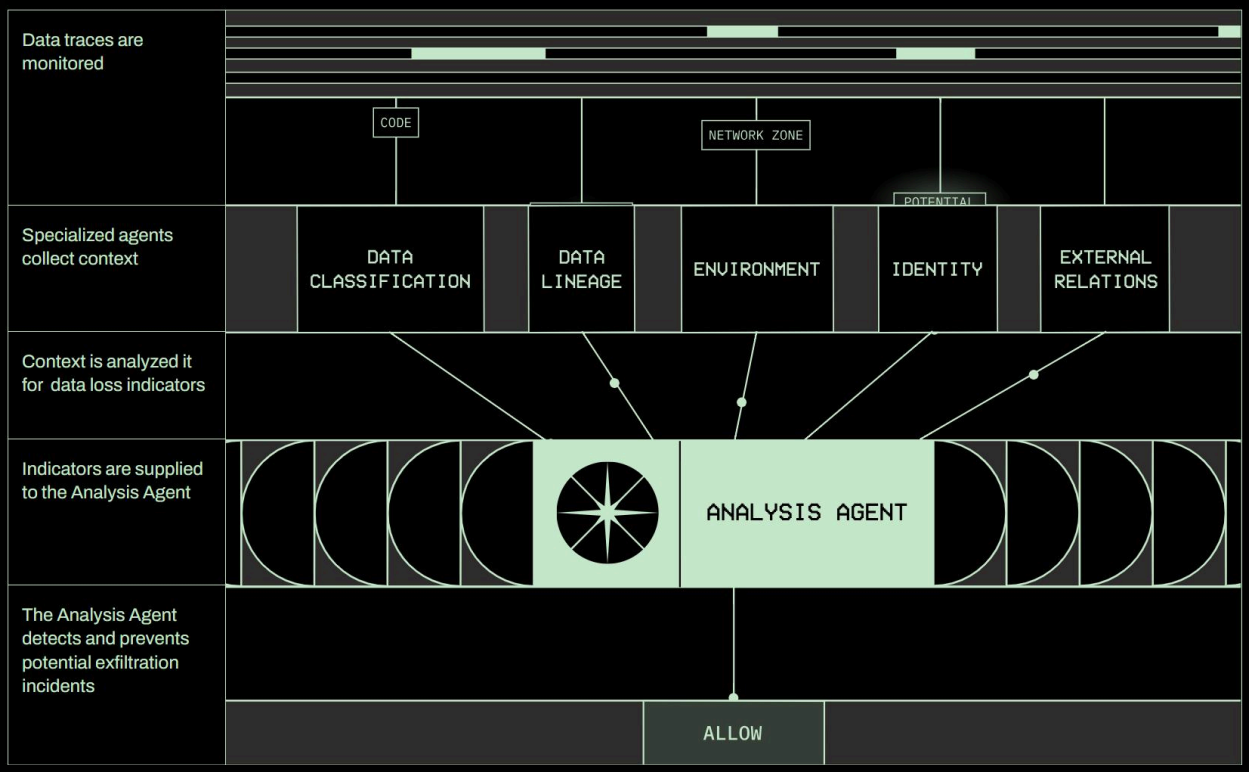
LEADERSHIP IMPERATIVES

Shadow AI is not a future problem; it is a present reality. Organizations that continue to rely on legacy DLP models and passive monitoring strategies will struggle to protect sensitive information in an AI-enabled world.

We must acknowledge that the use of AI is inevitable and design controls accordingly, with visibility, context, and real-time decisioning being foundational requirements. The goal is not to eliminate AI adoption, but to enable it safely. Success consists of embracing innovation, abandoning outdated assumptions, and deploying controls capable of operating at the speed and scale of modern AI-driven business.

Vendors today are developing context-aware proprietary AI agents that do more than just run regex rules based on fixed policy. These leading-edge solutions use dynamic reasoning to transform DLP from a policy-reliant system into an intelligent, context-aware one, while using policy-based enforcement for specific, deterministic use cases.

These tools can train on existing information flows to infer what is "normal" and then act in real time when detecting exfiltration events (inadvertent or deliberate). The scourge of business-interrupting false positives and SOC-fatiguing false negatives yields to a much more nuanced evaluation of data movement on endpoints, SaaS, email, storage, web, and more. By automatically classifying structured and unstructured data, these novel tools can map the flow of information through the organization, creating a data lineage that enables context-aware decision-making rather than policy-dependent decisions.



THE WAY FORWARD

AI-enabled enterprise is here to stay. We can no longer revert to a simpler time when basic tools and policy-based enforcement gave us "good enough" security. Compliance requirements such as GDPR and CCPA impose significant financial penalties on entities that fail to safeguard key data, with the largest fine in 2025 amounting to 1.2 billion euros (over $1.4 billion). **The cost of non-compliance exceeds any known tool or enforcement mechanism,** and as new jurisdictions add additional regulations, the risk of not protecting critical data becomes unacceptable. With stakes this high, the only logical approach is to use AI to contain AI.

Shadow AI represents a new type of data leakage risk that cannot be addressed by traditional monitoring techniques. It will remain a persistent challenge for security teams and represents a significant risk for organizations that continue to rely on legacy DLP solutions. An effective response is to deploy a set of solutions that autonomously identify, classify, track, and interdict information as needed.

The future is here, and so are the answers for those willing to research and innovate. Take the time to investigate the latest generation of solutions to protect and respond to data incidents at the speed of AI, so that you can offer your leadership the assurance that you are fielding the best protection available while avoiding compliance violations and fines that could wreck a business. Make data loss and non-compliance a thing of the past -- the solution is within your grasp.