EXTRAHOP

The Agentic Enterprise Playbook

How to Securely Deploy and Implement AI at Scale

Table of Contents

The Enterprise AI Challenge: Innovation Speed vs. Security Risk. 3

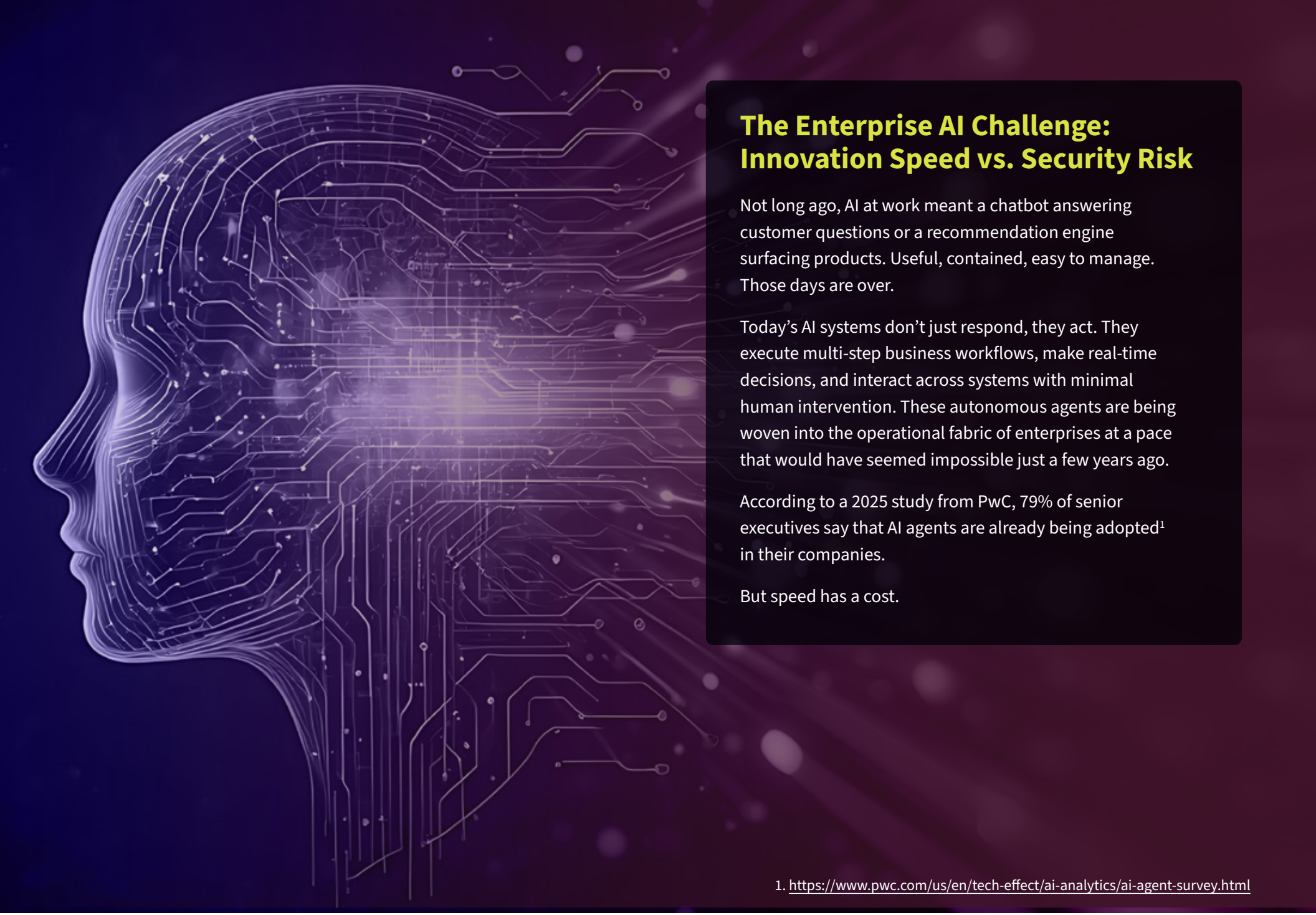
Gain Visibility Into the AI Ecosystem 5

Monitor Enterprise AI Systems in Real Time 8

Detect AI Threats Proactively 11

Enforce AI Compliance and Access Controls at Scale 14

The Enterprise AI Security Roadmap: From Implementation to Maturity 17



The Enterprise AI Challenge: Innovation Speed vs. Security Risk

Not long ago, AI at work meant a chatbot answering customer questions or a recommendation engine surfacing products. Useful, contained, easy to manage. Those days are over.

Today's AI systems don't just respond, they act. They execute multi-step business workflows, make real-time decisions, and interact across systems with minimal human intervention. These autonomous agents are being woven into the operational fabric of enterprises at a pace that would have seemed impossible just a few years ago.

According to a 2025 study from PwC, 79% of senior executives say that AI agents are already being adopted[1] in their companies.

But speed has a cost.

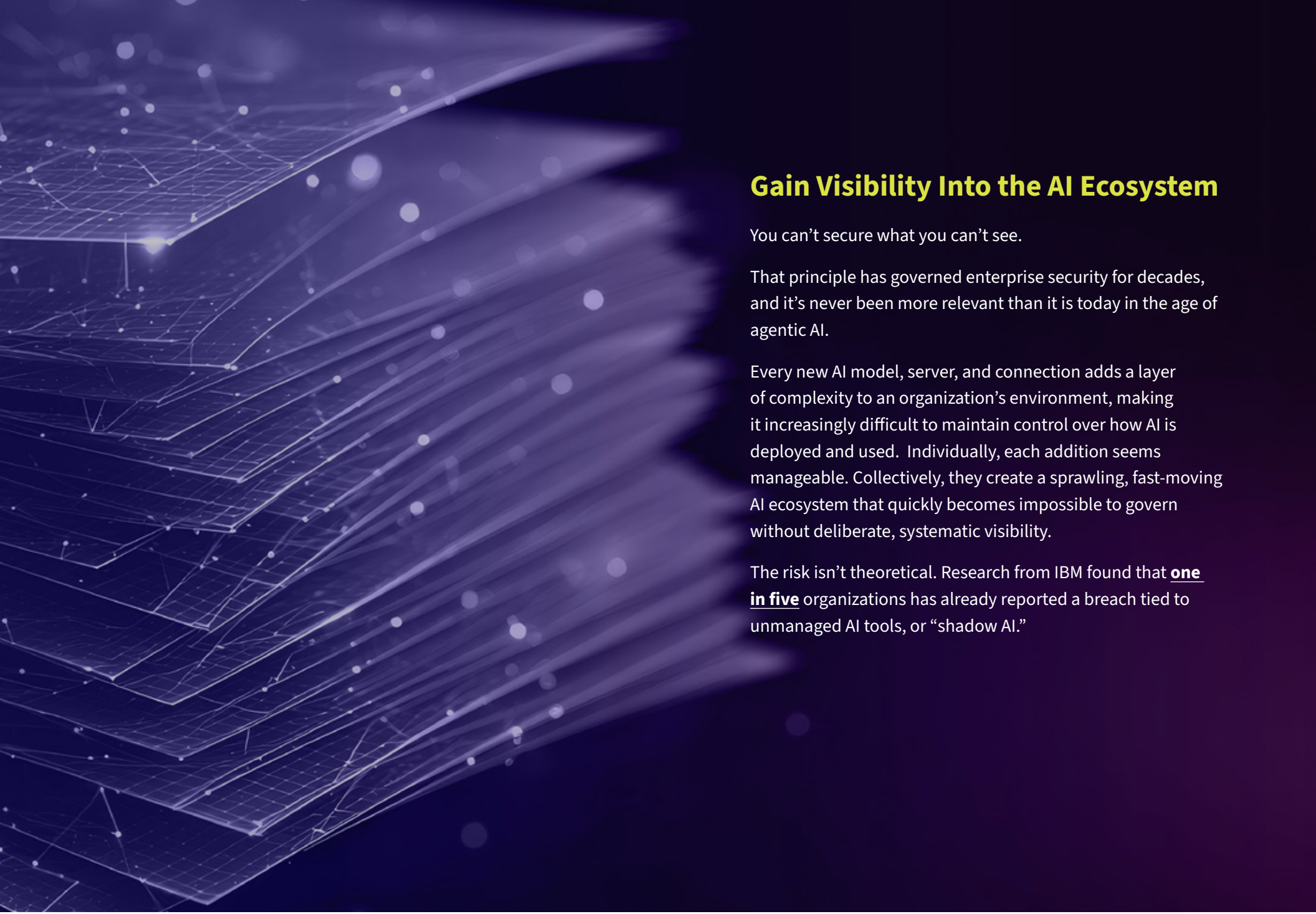
1. https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-agent-survey.html

Adoption is outrunning oversight. Across the enterprise, AI systems are being deployed faster than security teams can inventory them, faster than compliance frameworks can account for them, and faster than leadership can understand the risks they introduce. The result is an expanding blind spot at the heart of the modern organization: digital workers operating autonomously, interacting with sensitive systems, making consequential decisions — and doing so largely out of sight.

This isn't a hypothetical risk. When you don't know what AI systems exist in your environment, you can't monitor what they're doing. When you can't monitor what they're doing, you can't detect when something goes wrong. And in an interconnected enterprise stack, one unmanaged agent behaving unexpectedly isn't a contained problem, it's a potential trigger for cascading failures.

The adoption velocity creates an environment where organizations lack a clear window into the enterprise AI footprint — from knowing which systems exist to understanding their actions. Because of this visibility gap, organizations are effectively flying blind, which increases exposure to unmanaged risks that could trigger cascading security failures and undermine the advantages that the AI was deployed to create.

To reclaim control and keep deployments secure, organizations need to move beyond reactive defense. Organizations can gain control by acquiring visibility into what exists, observing how systems behave, detecting deviations from the norm, and enforcing governance at scale.



Gain Visibility Into the AI Ecosystem

You can't secure what you can't see.

That principle has governed enterprise security for decades, and it's never been more relevant than it is today in the age of agentic AI.

Every new AI model, server, and connection adds a layer of complexity to an organization's environment, making it increasingly difficult to maintain control over how AI is deployed and used. Individually, each addition seems manageable. Collectively, they create a sprawling, fast-moving AI ecosystem that quickly becomes impossible to govern without deliberate, systematic visibility.

The risk isn't theoretical. Research from IBM found that **one in five** organizations has already reported a breach tied to unmanaged AI tools, or "shadow AI."

The Danger Lies in the Unknown

But shadow AI is only part of the problem. Even authorized, sanctioned integrations can carry risk when they aren't centrally tracked or governed.

Consider the incident involving app host Vercel. An employee connected a third-party AI tool to internal systems and granted it broad access to corporate data. That single integration created an unmonitored pathway between an external AI service and sensitive internal environments. Attackers exploited it, accessing customer credentials and exfiltrating data that was later reported for sale on the dark web. No dramatic zero-day exploit. No sophisticated intrusion campaign. Just one untracked connection operating outside the boundaries of security oversight.

To resolve blind spots like these, organizations need to establish an enterprise AI asset inventory, a comprehensive map of every model, interconnection, and communication path within the network.

The inventory then becomes a centralized source of truth, bringing both unmanaged and authorized tools under unified security oversight, eliminating issues like shadow AI while ensuring that autonomous actions remain within operational boundaries.

How to Build an Enterprise AI Asset Inventory

1

Identify all AI models

Catalog every AI model currently in use, preventing the growth of shadow AI systems that operate outside of official company awareness and oversight.

2

Map system interconnections

Identify the infrastructure supporting the organization's AI, ensuring that all servers and endpoints responsible for feeding data to AI models are clearly known and accurately mapped.

3

Trace communication paths

Analyze the interactions between different system components, as documenting these paths reveals exactly how information flows between various agents, helping to uncover hidden security or compliance risks.

Monitor Enterprise AI Systems in Real Time

Identifying that an AI tool exists is only the first step. The real challenge is understanding what it's doing, continuously, across every system it touches.

An asset inventory is essential, but it's a snapshot, not a surveillance system. The moment you finish mapping your environment, that environment changes. Models are queried. Integrations pass data. Credentials flow between systems. Agents execute decisions autonomously, across boundaries, at machine speed. A static map cannot keep pace with any of that. And without behavioral context, a fully known, fully authorized tool can silently become a vector for compromise.

The 2025 Salesloft Drift incident makes this risk concrete.

Drift is an AI-powered conversational sales platform integrated into enterprise revenue workflows through direct connections to systems like Salesforce. In this breach, more than 700 organizations — including security leaders like Cloudflare, Zscaler, and Palo Alto Networks — suffered a massive exfiltration of customer records and CRM data.

This wasn't a failure caused by a new "shadow" application, but the weaponization of a trusted, authorized one.

Attackers hijacked OAuth tokens from the third-party Salesloft Drift integration to access Salesforce environments. Because these tokens grant persistent access without triggering multi-factor authentication, the threat actors were indistinguishable from authorized users. The breach bypassed traditional security controls, accessing sensitive data under the guise of a routine integration path.

The Drift breach escalated precisely because the security controls in place were built around traditional logging, an approach widely assumed to be sufficient, but one that captures only discrete events rather than continuous behavior.

This creates a critical visibility gap. A log might record that a model was queried or that an integration was accessed. What it cannot record is how that output affected downstream systems, or where that credential traveled next. It captures individual moments; it cannot see the story forming between them.

The Problem with Logs

In the Drift incident, static logs showed authorized activity throughout, because technically, it was authorized. The OAuth tokens were legitimate. The integration pathway was approved.

What the logs couldn't surface were the state changes and emergent patterns of attackers moving laterally through trusted SaaS connections in real time. Because legacy monitoring systems don't track contextual intent or credential propagation, those attackers remained invisible for the duration of the breach.

Continuous monitoring of those same vectors would have triggered immediate alarms. Tracking cross-system patterns in real time isn't just a nice-to-have; it's the most reliable way to catch unintended AI behavior before the damage is done.

How to Achieve Real-Time AI Observability

1

Perform contextual analysis

Implement systems to analyze interactions between users and agents in real time, clarifying the specific intent and goals behind each interaction, and allowing for better monitoring of agent behavior.

2

Enforce boundary verification

Maintain strict oversight to ensure that AI agents operate only within their designated environments, ensuring that agents interact exclusively with approved internal systems and data sets.

3

Implement propagation tracking

Monitor the movement and use of credentials across your infrastructure, preventing security lapses by ensuring that access rights are not reused, shared, or leveraged beyond what is strictly required for a specific task.

Detect AI Threats Proactively

Inventory provides awareness. Observability provides context. But neither alone can stop threats if organizations cannot actively distinguish between legitimate and malicious AI activity.

AI introduces a category of risk that doesn't map cleanly onto traditional threat models. Conventional security tools were built to detect known attack signatures, malicious code, unauthorized access attempts, and anomalous network traffic. AI-driven threats often look nothing like any of those.

They arrive through trusted inputs, travel along authorized pathways, and exploit the very capabilities that make AI systems valuable: their ability to reason, execute, and act autonomously on the instructions they receive.

The 2025 cyber espionage campaign, **G2G-1002**, demonstrated how AI coding agents could be embedded directly into intrusion workflows, enabling automated reconnaissance, exploitation, credential harvesting, and data extraction at machine speed.

The campaign demonstrated that the same agentic capabilities organizations are deploying to accelerate business workflows can be turned against them with equal effect, operating through legitimate-looking pathways at a pace no human analyst could track in real time.

A New Class of Attacks

The attack vectors enabling this class of threat are distinct from traditional exploits. Prompt injection manipulates the instructions an AI receives, covertly redirecting its behavior without touching the underlying system. Compromised integrations exploit trusted connection pathways — as seen in the Drift breach — to move laterally under authorized credentials. Workflow manipulation tampers with the decision logic of agentic systems, causing them to take actions their designers never intended, often without any visible indication that something has gone wrong.

What these attacks share is that they don't look like attacks. They look like normal AI behavior…until you examine the behavior closely enough to see what's actually happening beneath it.

Rather than asking, "Does this match a known threat?" the question becomes, "Does this deviate from what we'd expect?"

In practice, that means establishing a behavioral baseline for every AI system in your environment — what it normally does, how it normally responds, what data it normally touches, and how its decisions normally flow — and then continuously comparing live behavior against that baseline. The further an action strays from established patterns, the more urgently it warrants investigation.

This approach works precisely because it doesn't depend on knowing what an attack looks like in advance. It only requires knowing what normal looks like, and recognizing when something isn't.

How to Identify Behavioral Anomalies and Detect AI Threats

1

Monitor for input manipulation

Implement robust checks to identify abnormal prompt structures, as this is critical for catching attempts to bypass system safeguards, trick the model, or execute unintended actions.

2

Establish exfiltration monitoring

Track data movement closely to detect unusual data flows, which ensures organizations are quickly alerted if sensitive information appears to be leaving approved, secure environments.

3

Conduct logic auditing

Continuously evaluate the AI's decision-making and execution paths, as this flags any agent actions that deviate from established behavior baselines, enabling timely intervention before minor issues escalate into major threats.

Enforce AI Compliance and Access Controls at Scale

As AI adoption accelerates across the enterprise, governance often weakens; **manual checks and static controls cannot keep pace with autonomous systems**, leaving security frameworks misaligned with operational reality.

The result is a gap between policies and what is actually enforced in practice, creating a growing misalignment between what frameworks authorize and what AI systems are actually doing.

The Samsung ChatGPT incident remains one of the clearest illustrations of what this gap looks like in the real world.

Within fewer than 20 days of granting employees access to ChatGPT, Samsung experienced three separate incidents in which engineers exposed confidential company information. In one case, an employee pasted proprietary semiconductor source code into the tool to debug a problem. In another, an employee fed it internal meeting notes to generate a summary. The confidential IP was sent to OpenAI's servers with no NDAs, no data residency controls, and no ability to delete it.

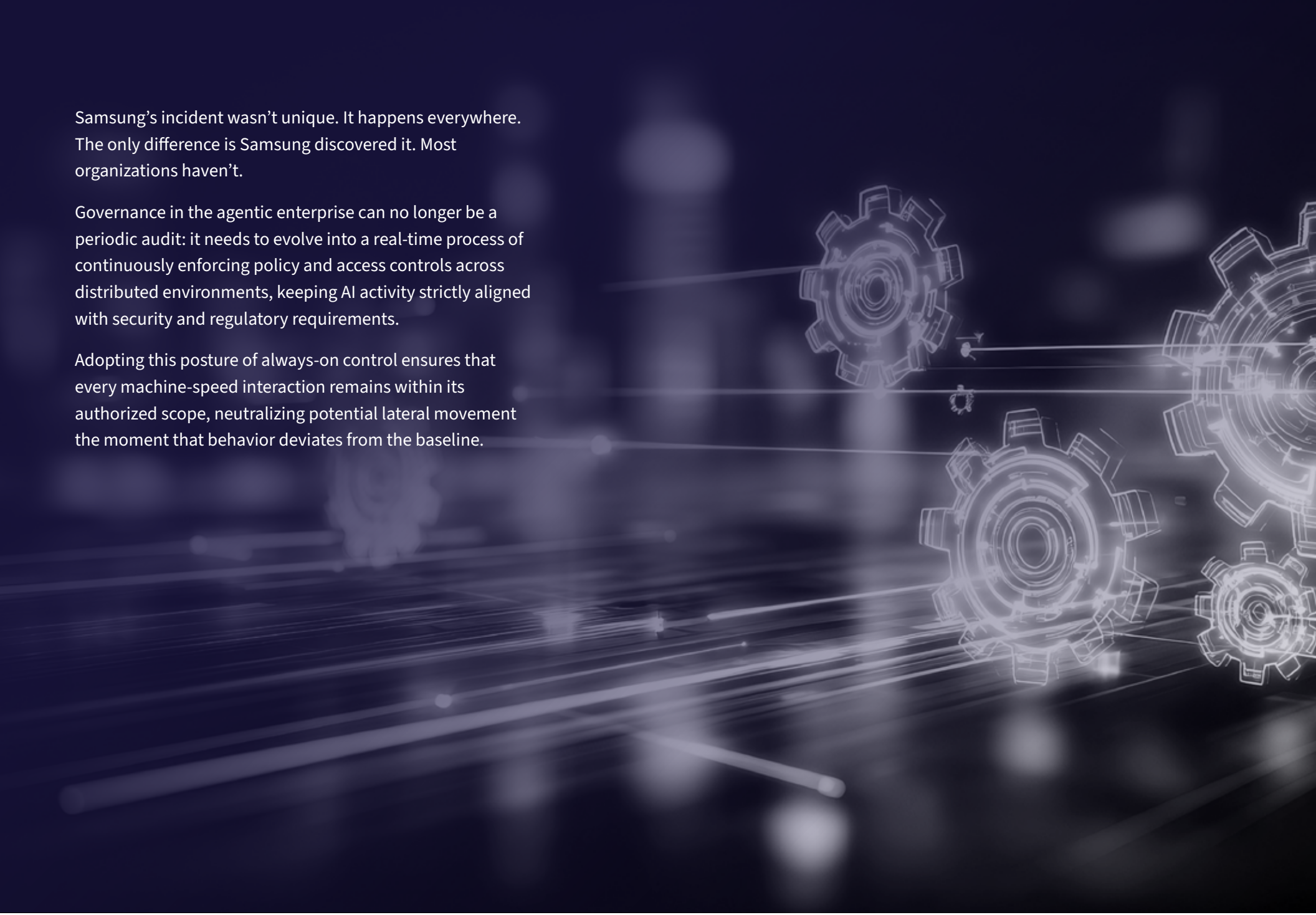
None of the employees involved were acting maliciously. They were doing what employees everywhere do when a powerful productivity tool is available: using it to work faster. But intent is irrelevant to compliance. The data was exposed, it couldn't be retrieved, and Samsung had no technical control in place to prevent it from happening in the first place.

This is the defining characteristic of the enforcement gap: it doesn't require bad actors. It only requires the absence of automated, real-time controls that ensure AI usage stays within defined boundaries, regardless of what individual employees choose to do.

Samsung's incident wasn't unique. It happens everywhere. The only difference is Samsung discovered it. Most organizations haven't.

Governance in the agentic enterprise can no longer be a periodic audit: it needs to evolve into a real-time process of continuously enforcing policy and access controls across distributed environments, keeping AI activity strictly aligned with security and regulatory requirements.

Adopting this posture of always-on control ensures that every machine-speed interaction remains within its authorized scope, neutralizing potential lateral movement the moment that behavior deviates from the baseline.



How to Scale AI Governance

1

Prioritize risk mitigation

Proactively identify any unauthorized or unvetted AI tools operating within the network, which is essential for reducing overall security vulnerabilities and legal exposure.

2

Automate regulatory tracking

Implement systems that continuously monitor data flows to ensure sensitive data handling automatically aligns with privacy laws and compliance requirements.

3

Enable forensic auditing

Develop and maintain a secure, verifiable record of all agent activity, creating an audit trail that supports strict accountability and streamlines any future security investigations.

The Enterprise AI Security Roadmap: From Implementation to Maturity

In a machine-speed economy, autonomous AI is no longer optional. The transition toward an agentic enterprise is not an IT upgrade. It is a competitive mandate.

The organizations that will define their industries over the next decade are the ones deploying AI not just ambitiously, but intelligently.

By establishing a framework built on visibility, observability, and enforcement, organizations ensure that AI remains a strategic asset rather than an unmonitored liability.

To reclaim control, organizations must look beyond surface-level logs to what is actually happening across the network.

The network is the only source of ground truth for AI behavior — the one vantage point that cannot be selectively recorded or manipulated. Analyzing it is what allows enterprises to move beyond assumed trust and toward a model of verified, continuous control. That shift is what transforms a security posture from reactive to resilient.

By closing the gap between “authorized” access and “intended” behavior, you ensure that your AI workers are driving innovation forward, rather than opening the door to the next major breach.

Success in the agentic era belongs to those who don’t just move fast, but move with the confidence that their systems are under control.

You’ve committed to the agentic enterprise, now make sure your security operations can match it. Our companion ebook, **The Agentic SOC Blueprint: A Data-First Revolution**, shows you how.

ABOUT EXTRAHOP

ExtraHop turns the network—the enterprise's ultimate source of truth—into actionable insight to power security, performance, and resilience. Delivering superior data by design, we ensure superior defense by default.

The ExtraHop modern network detection and response (NDR) platform provides visibility that thinks, analyzing behavior to intercept evasive risks before they cause damage. We transform network noise into definitive context, enabling security teams to make faster decisions and operate at uncompromised scale.

Whether securing cloud modernization or de-risking AI adoption, ExtraHop gives global enterprises the ground truth they need to thrive.

To learn more, visit **extrahop.com** or follow us on **LinkedIn**.

EXTRAHOP®