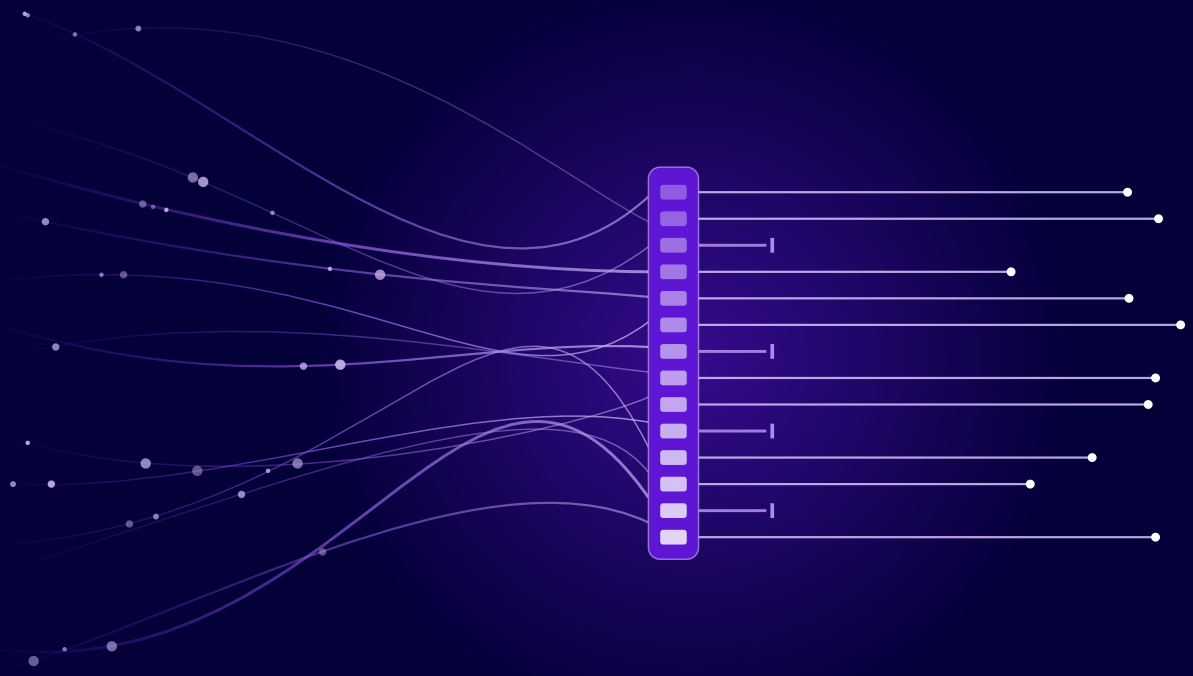
Nightfall AI

You enabled AI. *Now what?*

What the data says about how AI moves enterprise data, what has already gone wrong, and how to keep the productivity without inheriting the exposure.



AI moves your data. Nightfall controls it.

NIGHTFALL.AI

01 · EXECUTIVE SUMMARY

Adoption outran *control.*

Nearly nine in ten companies now use AI in at least one business function, and most of them turned it on faster than they could govern it. The gap shows up in the breach data.

43%
of breaches in 2026 involved shadow AI, up from 20% the year before

92%
of organizations with an AI-related breach lacked proper AI access controls

$4.99M
global average cost of a data breach in 2026, up 12% in a single year

Source: IBM Cost of a Data Breach Report 2026.

The risk is not that AI exists. It is that AI moves data.

Prompts carry source code and contracts. Copilots read every file a user can reach. Coding agents run shell commands and call tools. MCP servers connect models to databases, ticketing systems and email.

Each of these is a new path for sensitive data to leave the company, and most of them sit outside the controls built for the last decade of SaaS and email.

What this paper covers

What the data says about how AI is being used, where sensitive data is actually going, what has already gone wrong, and what regulators now expect.

It then describes a control framework that lets a company keep the productivity it bought and close the exposure it inherited, and shows how Nightfall delivers that framework across humans and AI agents on every surface.

> *The companies getting breached through AI are, in large part, the companies that never wrote down the rules.*

IN THIS PAPER

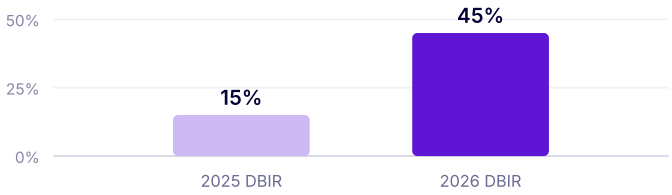
02 The adoption reality
03 Where the data actually goes
04 What has already gone wrong
05 The regulatory clock
06 Why the old playbook fails
07 What good looks like
08 How Nightfall does it
09 A 90-day plan

02 · THE ADOPTION REALITY

Pilot to default in *three years.*

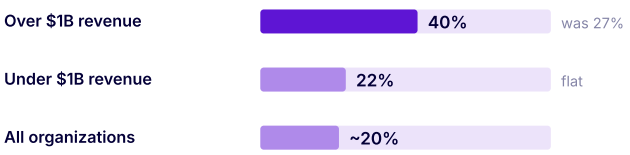
McKinsey's 2026 global survey of 1,719 organizations found nearly nine in ten using AI regularly in at least one business function, 56% using it in three or more, and 44% reporting that AI is scaling across the enterprise.

Employees who are regular users of AI tools at work



Source: Verizon Data Breach Investigations Report 2026.

Organizations scaling AI agents, by company size



Source: McKinsey, The state of AI in 2026 (fieldwork May to June 2026).

Agents are following the same curve

Among companies with more than $1 billion in revenue, 40% are scaling AI agents, up from 27% the year before. These are not chatbots answering questions. They are systems that read files, call tools and act on data with limited human review.

Employees moved faster than their employers

Verizon found that 67% of users accessing AI services on corporate devices were doing so through non-corporate accounts, and that source code is the number one data type uploaded to unauthorized AI services. A personal ChatGPT or Gemini login on a work laptop is now the common case, not the exception.

Governance did not keep up

In IBM's 2025 study, 63% of organizations had no AI governance policy at all. In the 2026 study, 68% of breached organizations still lacked AI governance to manage AI or detect shadow AI.

> *67% of corporate-device AI access ran through accounts the company could not see.*
>
> VERIZON DBIR 2026

03 · WHERE THE DATA ACTUALLY GOES

Six surfaces, *mostly ungoverned.*

Sensitive data leaves through six kinds of AI surface. Most companies have controls for, at best, one of them. The common thread: each channel was designed for capability, not for control.

Prompts and uploads
Chat tools in the browser. Source code, contracts and customer records pasted in or attached, often from personal accounts.

Copilots inside SaaS
Microsoft 365 Copilot, Gemini for Workspace. They read whatever the user can read, including over-shared drives and channels.

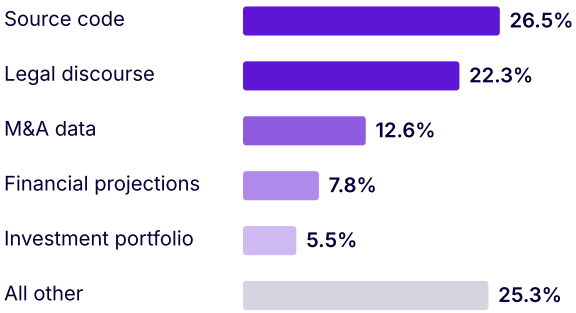
Coding agents and CLIs
Claude Code, Cursor, Codex, Copilot, Kiro. They read repos, run shell commands and write to disk, including secrets.

MCP servers
Local and remote tool connections that let agents query databases, read email and open tickets. Usually installed with no review.

Managed and homegrown agents
Copilot Studio, Agentforce, internal agents on LiteLLM or a gateway. Standing credentials to CRM, HR and finance.

Desktop AI apps
Claude, ChatGPT, Ollama. They run outside the browser, where browser-based controls and CASB proxies cannot see them.

What sensitive data looks like inside enterprise prompts



Share of 579,113 sensitive-data instances found in 22.5 million enterprise prompts and uploads, calendar 2025. Nearly 6,000 were live credentials.

The coding agent is not malicious. It is fast.

GitGuardian counted 28.65 million new hardcoded secrets on public GitHub in 2025, up 34% in a year, with AI-service secrets up 81%. Commits assisted by one leading coding agent showed a 3.2% secret-leak rate against a 1.5% baseline.

The agent does what it is told, including pasting an API key into a file. The question for every surface is the same: what data is in this flow, who or what is moving it, and should it be allowed?

> *MCP was built for interoperability, ease of use and flexibility first.*
>
> AARON LORD, GARTNER, APRIL 2026

04 · WHAT HAS ALREADY GONE WRONG

Measurable, and *more expensive.*

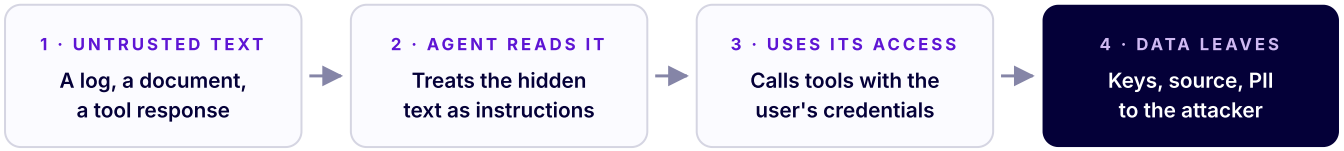
21%
of all breaches in 2026 were AI-related, up from 13% a year earlier

$5.39M
average cost of a breach involving shadow AI, up from $4.63M a year earlier and above the $4.99M global average

247
days, on average, to identify and contain a breach

Source: IBM Cost of a Data Breach Report 2026. AI-driven attacks added about $1 million per breach above the malicious-breach average.

How an agent becomes the exfiltration path



Documented incidents and disclosed flaws, past twelve months

OCT 2026
GitLab AI Gateway, CVSS 9.9. An authenticated user could escape the prompt-template sandbox and run commands on the gateway.

AUG 2026
"Ghostjacking" against Claude Code. Poisoned log data hijacked the agent in 9 of 10 tests; 2,700+ exposed Datadog keys found along the way.

AUG 2026
GhostSplice MCP research. A refused exfiltration request, split across MCP channels, went from 0% to 100% compliance on three major models in lab tests.

JUL 2026
Copilot for Word worm. Hidden instructions altered generated figures and copied themselves into every new document built from the source. Disclosed after 144 days.

MAY 2026
Composio credential theft. One employee's Gmail OAuth token led to roughly 10,200 customer credentials exposed, including 5,001 GitHub tokens.

APR 2026
MCP STDIO design flaw. Remote code execution exposure across 200,000+ MCP servers and 150 million SDK downloads.

SEP TO NOV 2025
AI-orchestrated espionage campaign. Roughly 30 organizations targeted; AI executed an estimated 80 to 90% of the operation.

Three lessons underneath the list

The attacker does not need to break in. They need to get text in front of an agent that already has access.

Credentials are the currency. One token, one exposed key, one over-scoped OAuth grant becomes thousands of downstream doors.

The agent and the insider are now the same problem. An agent acting on a poisoned instruction looks, from the data's point of view, exactly like an employee exfiltrating on purpose.

Gartner expects it to get worse first

By 2028, 25% of enterprise generative AI applications will suffer at least five minor security incidents a year, up from 9% in 2025. AI-driven applications will account for half of all incident response work by the same year.

05 · THE REGULATORY CLOCK

Regulators stopped *waiting.*

Nearly every obligation below comes back to two questions: what did AI access, and what did it do with it?

JURISDICTION	RULE	STATUS	WHAT IT REQUIRES
EU	AI Act, Article 50 transparency	In force since Aug 2, 2026; watermark grace to Dec 2, 2026	Disclose AI interaction, label AI-generated content. GPAI fines up to 15M euros or 3% of global turnover.
EU	AI Act, high-risk systems (hiring, credit)	Dec 2, 2027 (delayed by Digital Omnibus)	Risk management, logging, human oversight, data governance.
EU	AI Act, high-risk in regulated products	Aug 2, 2028	Same obligations for AI in medical devices, machinery and similar.
US, Texas	TRAIGA	In force since Jan 1, 2026	Intent-based liability. $10,000 to $200,000 per violation, 60-day cure.
US, California	SB 53 and AB 2013	In force since Jan 1, 2026	Frontier model safety frameworks, incident reporting, training data disclosure.
US, California	AI Transparency Act (SB 942)	Operative since Aug 2, 2026	Latent disclosures and a public detection tool for generative AI providers. $5,000 per violation, each day a separate violation.
US, California	CCPA automated decision-making rules	Jan 1, 2027	Notice, opt-out and risk assessment for significant decisions.
US, Colorado	SB 26-189 (replaces the 2024 AI Act)	Jan 1, 2027 (signed May 14, 2026)	Pre-use notice and 30-day explanations for consequential decisions. Attorney general enforcement, 60-day cure.
US, Illinois	Human Rights Act amendment	In force since Jan 1, 2026	AI in hiring and promotion liable for discriminatory effect, no intent required.

The federal picture is unsettled

A December 2025 executive order directed the Justice Department to challenge state AI laws, and the DOJ joined xAI's suit against Colorado in April 2026. No federal preemption statute exists, so companies operating across states carry the strictest applicable rule.

Sector regulators applied the old rules

HIPAA does not care whether PHI left through email or through a copilot. PCI DSS does not distinguish a cardholder number pasted into a ticket from one pasted into a prompt. SOC 2 auditors now ask for AI access logs alongside everything else.

> *Can you show me what the AI touched last quarter, and prove we stopped what it should not have?*
>
> THE AUDIT COMMITTEE QUESTION, 2026

Gartner projects that through 2027, manual AI compliance processes will expose 75% of regulated organizations to fines above 5% of global revenue.

06 · WHY THE OLD PLAYBOOK FAILS

Three responses, *three failures.*

Most companies have tried one of these. The underlying mistake in all three is treating AI as a new application to be allowed or denied, rather than as a new actor that moves data.

1

Blocking the tools

Banning ChatGPT at the firewall pushed usage to personal phones and personal accounts, which is exactly where Verizon found two thirds of corporate-device AI access happening.

It also forfeits the productivity the company paid for, and does nothing about the copilot already licensed inside Microsoft 365 or the coding agent already on every engineer's laptop.

2

Extending legacy DLP

Legacy data loss prevention was built around static rules, known channels and human behavior: a regex for a card number, applied to email and file shares.

It was never designed to inspect a tool call from an agent, a streamed model response, or a shell command. Browser controls miss desktop apps. CASB proxies miss local MCP. Endpoint detection sees a network call, not the Social Security number inside it.

3

Buying a dashboard

A new category of AI governance tools promises visibility: which AI apps employees use, which agents exist, how much traffic flows.

Visibility is necessary and not sufficient. Knowing a developer connected an agent to the production database at 2:14 p.m. is useful only if something could have stopped the query that followed.

> *Visibility without control is just a dashboard.*

The frame has to change

THE OLD UNIT OF RISK



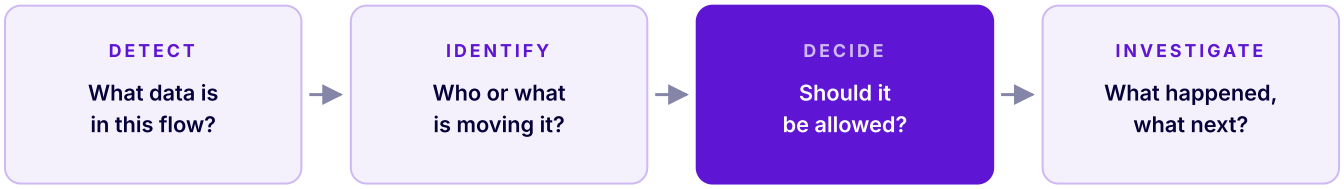
THE NEW UNIT OF RISK: THE WORKFLOW



07 · WHAT GOOD LOOKS LIKE

Four questions, *answered in real time.*

A company has AI under control when it can answer these for every surface where AI touches data, at the moment the data moves.



QUESTION	WHAT IT REQUIRES	WHAT IT RULES OUT
What data is in this flow?	Detection that understands content: PII, PHI, payment data, credentials, source code and corporate IP, inside prompts, files, tool calls, tool responses, shell commands and model output.	Pattern matching alone. Inspection limited to the browser.
Who or what is moving it?	An inventory of AI apps, agents and MCP servers, each tied to a human identity and an agent identity.	Anonymous agent traffic. Unknown local MCP servers.
Should it be allowed?	Policy aware of the destination: the same document is fine in an approved enterprise tool and not fine in a personal account or an unreviewed MCP server.	One rule for all AI. Allow-or-deny at the app level.
What happened, and what next?	Lineage from origin to destination across human and agent activity, with every event turned into an investigation a security team can act on.	Raw alert streams. Logs with no context.

Enforce at the moment of movement

Before the data leaves, with a response proportionate to the risk: block, redact, coach the user, or ask for approval. Post-hoc logging is evidence for the breach report, not a control.

Two actors, one policy

An employee pasting a customer list into a chatbot and an agent reading that list through an MCP tool are the same exposure. A company should not need two products and two rulebooks to govern them.

> *Security teams are not asking for a kill switch. They want to say yes, with conditions.*

WHAT WE HEAR IN THE FIELD

08 · HOW NIGHTFALL DOES IT

AI moves your data. *Nightfall controls it.*

Nightfall is the AI-native data security platform that gives enterprises real-time control over sensitive data movement across humans and AI agents, MCP servers, SaaS, email and endpoints. One platform. Every surface. Every actor.



Detection depth

AI-native detection built on supervised fine-tuned models, not regex, finds sensitive data and corporate IP in prompts, uploaded files, tool calls and responses, shell commands and model output. It understands transformed and AI-generated content. More than 130 detectors cover PII, PHI, payment data, credentials, source code and IP.

Remediation breadth

Detection triggers a response before the data leaves: block, redact and let the rest through, coach the user, or route for approval. Rules are aware of the destination. Lineage traces each item from origin to destination across human and agent activity.

Agent intent and identity

A catalog of every agent and MCP server in use, including local servers on developer laptops that cloud gateways cannot see, enriched with agent identity. Nightfall flags agent sessions and prompts that drift from the task the agent was given. Enforcement is synchronous: the agent cannot proceed until Nightfall answers.

Insider risk for people and agents

Every event ties to both the person's and the agent's identity, correlated across endpoint, SaaS and AI/MCP. Autonomous investigation turns each event into CISO-ready context and a recommended action. Nightfall connects by API where the data lives, with a lightweight endpoint agent and native hooks for coding agents and local MCP. Observe first, enforce surface by surface.

> *Legacy DLP was not built for AI. Nightfall was.*

09 · A 90-DAY PLAN

From enabled to *governed.*

This does not require a multi-year program. Each step below produces something a board or auditor can see.

DAYS 1 TO 30	DAYS 31 TO 60	DAYS 61 TO 90
See what is moving	Write the rules, turn on controls	Close the loop
Inventory · Observe	Policy · Enforce · Approved MCP registry	Agents · One queue · Audit report

1

See what is moving

- Inventory every AI app, agent and MCP server in use, including personal accounts on corporate devices and local servers on developer machines.
- Run detection in observe mode across prompts, uploads, tool calls and copilots.
- Expect to find sensitive data in AI flows within the first week, usually source code and customer records in personal accounts.

2

Write the rules, turn on controls

- Define which data classes may go to which destinations.
- Enforce where exposure is highest and business impact lowest: block credentials and PHI outright, redact PII headed to unapproved tools, coach users on the rest.
- Approve a registry of sanctioned MCP servers and flag everything else.

3

Close the loop

- Extend enforcement to coding agents and managed cloud agents.
- Tie every agent to a human owner. Route investigations through one queue for human and agent events.
- Produce the first AI access report for the audit committee: what AI touched, what was stopped, what changed.

The outcome is a company that can keep expanding AI use because it can prove, at any moment, *what the AI did with its data.*

See it. Understand it. Stop it before it leaves.
Nightfall secures the first data flows in minutes, not quarters. Start in observe mode and turn on enforcement surface by surface.

nightfall.ai

SOURCES

Where the *numbers come from.*

McKinsey & Company. The state of AI in 2026. Survey of 1,719 participants in 97 nations, May 4 to June 8, 2026. mckinsey.com

IBM. Cost of a Data Breach Report 2025 and 2026, "The AI tipping point." Research conducted with the Ponemon Institute. ibm.com/reports/data-breach

Verizon. 2026 Data Breach Investigations Report. verizon.com/dbir. Figures as reported in the DBIR and its published summaries.

Industry research. Analysis of 22.5 million enterprise GenAI prompts and file uploads across 665 AI tools, January 1 to December 31, 2025.

GitGuardian. The State of Secrets Sprawl 2026. gitguardian.com

Gartner. "Gartner Predicts AI Applications Will Drive 50% of Cybersecurity Incident Response Efforts by 2028," March 17, 2026. "Gartner Predicts 25% of All Enterprise Gen AI Applications Will Experience at Least Five Minor Security Incidents Per Year by 2028," April 9, 2026. gartner.com

Incident timeline. awesome-ai-agent-attacks, a sourced public timeline of AI agent security incidents 2024 to 2026, with primary reporting from Anthropic, The Hacker News, The Register, BleepingComputer, Infosecurity Magazine, Cymulate, OX Security, Kaspersky, Material Security and GitLab.

Regulation. "AI Regulation in August 2026: What Is Actually in Force," Ztek Cyber. Primary texts: EU AI Act and Digital Omnibus on AI; Texas TRAIGA; California SB 53, AB 2013, AI Transparency Act and CCPA ADMT regulations; Colorado SB 26-189; Illinois Human Rights Act.

Nightfall AI. Product and platform descriptions reflect Nightfall's current platform and messaging as of November 2026.

© 2026 Nightfall AI. All rights reserved. Third-party names and marks are the property of their respective owners. Statistics are reported as published by the cited sources and have not been independently verified by Nightfall.