You Can't Secure the AI You Can't See

Why safe AI adoption starts with knowing what's already running in your business

Not long ago, AI at work meant one thing: an employee opened a browser tab, typed a question into a chatbot, and read the answer. That is no longer true. AI now runs as software installed on laptops. It writes and executes code. It reads files, drafts email, queries databases, and updates records inside the systems the business depends on. And it is very often reached through an employee's personal account rather than a company one.

> Verizon's 2026 Data Breach Investigations Report found the share of employees regularly using AI on company devices rose from 15% to 45% in a single year. 67% of those employees used personal accounts rather than company-managed ones.

This is not a story about a security program that failed. Nothing failed. A category of technology arrived faster than the controls designed to govern it, and it arrived everywhere at once. Well-run organizations with capable IT partners are in precisely the same position.

Written AI policy and actual AI behavior have drifted apart. Most organizations cannot say, with evidence, which AI tools run on their computers, which accounts those tools sign in with, or what actions they are permitted to take. A policy describing a reality nobody has measured is not governance. It is a hope.

This is a measurable problem. AI leaves a trail on the devices where it runs. An **AI Risk Assessment** observes what is actually running across company-managed computers and replaces assumption with evidence. Evidence first. Then decisions. Then control.

AI in 2024

- One browser tab, one chatbot
- Answered questions
- A single approved tool
- A written policy could cover it

AI in 2026

- Desktop apps, coding assistants, agents
- Takes action inside business systems
- Personal accounts, unreviewed tools
- Policy alone cannot see it, or stop it

What Actually Changed

Four shifts happened at once. Together they moved AI beyond the reach of the controls built to govern software.

AI moved onto the device. The browser is now the smallest surface. AI runs as installed desktop software like ChatGPT and Claude, as coding assistants like GitHub Copilot, Cursor, and Claude Code, as command-line agents, and as local models that never touch the internet.

Identity became ambiguous. A consumer plan and an enterprise plan look identical on screen while offering entirely different administrative visibility, retention settings, and authentication requirements. Sign into a personal AI account on a company laptop and the organization cannot see that session or audit what moved through it.

AI started taking action. Modern AI does not just produce text. It modifies source code, executes commands, and completes multi-step tasks inside business systems. Gartner projects that by the end of 2026, 40% of enterprise applications will include task-specific AI agents, up from fewer than 5% in 2025. A tool that answers a question is a data risk. A tool that can delete a record is an operational one.

Connectors opened everything else. Connectors let an AI reach into an outside system: a code repository, a database, a customer record. Some are official. Some are written by an employee in an afternoon. Each carries its own credentials, permissions, and ability to read, write, or delete.

Where AI runs	What it can do	Why it is hard to see
Web browser	Chat, upload files, run extensions	Looks like ordinary web traffic
Desktop applications	Reach local files and software	No website, so no domain to govern
Coding assistants	Modify code, execute commands	Looks like ordinary developer work
Connectors and integrations	Reach email, code, cloud, CRM	Configured per user, never inventoried

What an AI Risk Assessment Finds

An assessment answers three questions, in order. **What is actually running. What it means. What should change.** Every finding belongs to a real tool, on a real machine, used by a real person.

1

Discover

What AI is running

2

Understand

What the risk is

3

Decide

What is permitted

4

Control

Enforce and monitor

What is running. Every AI tool present, and how it runs: in the browser, as a desktop app, as a coding assistant, or as a local model. Which tools are reached through personal rather than company accounts. Which sit on consumer plans with no administrative controls. Where multi-factor authentication is off, and where settings permit conversations to be used for training.

Where sensitive information appears. Credentials and personal data surfacing inside AI conversations, on the tools where that detection is supported.

What the agents can reach. Which AI tools take action, and what they touch. Findings commonly include credentials stored in plain text, unencrypted connections, integrations permitted to delete records, and destructive commands available to an agent.

What it does not do. An assessment observes. Nothing is blocked, nothing is changed, and no employee is interrupted. It runs on company-managed computers only. Prompt content is not stored by default, and detailed logging stays off unless deliberately enabled. An initial inventory arrives quickly, and observation continues for a period your IT provider defines.

What comes after. The assessment is a beginning, not the destination. Once leadership decides what should be permitted, controls follow: warn before a risky action, restrict personal-account sign-in, prevent a destructive operation before it executes. Visibility, then governance, then control, continuously.

Under the Hood

Where the visibility comes from. A lightweight component runs on the endpoint, paired with a managed extension for Chromium browsers. It covers Windows and macOS workstations, and installs through the endpoint management tooling an organization already operates.

Observation happens on the device, not the network, because the context does not exist on the wire. Conventional tooling sees network events, processes, and file writes. It cannot see the prompt, which integration was invoked, or whether a human or an agent initiated the action. Many integrations use the Model Context Protocol, or MCP, the standard for connecting AI to outside systems. **This is additive.** It sits alongside endpoint detection, data loss prevention, and cloud access controls, and replaces none of them.

How a risk becomes a finding. Policy sets a target score for each risk attribute. The endpoint scores what actually exists. When observed exceeds target, an issue is raised against a specific device and user. Severity runs 0 to 39 low, 40 to 59 medium, 60 to 79 high, above 80 critical.

Evaluated on AI applications	Evaluated on connectors and integrations
Personal versus company account	Local command execution
Sensitive data handling	Plain-text credentials
Authentication and training-data settings	Permission scope and enforcement
Connectors and model in use	Unencrypted and destructive operations

Control operates on the operation, not the tool. Switching an integration off removes the productivity that caused someone to adopt it. A delete operation inside a specific tool call can be disallowed while every other operation keeps working.

Data handling. Prompt content is not retained by default, and audit logging is opt-in and redacted. Discovered credentials are not uploaded. Findings export as a report, a device list, or an API.

Ask your IT provider about performing an AI Risk Assessment. Find out what AI is actually running in your business, what it can reach, and what it is doing.

Learn more at **https://greenlightcyber.com/mcs/ai-security**