

The AI Governance Playbook

How to Adopt AI Safely in 2026

Adoption has already happened. This is how security catches up — the visibility, controls, and policy framework that let your people work with AI confidently, and let your security team see what they're defending.

INTRODUCTION

AI is the engine of transformation — and a fast-growing source of risk

AI has become the engine of business transformation. It is also one of the fastest-growing sources of security risk. Employees use AI tools every day to draft content, analyse data, and automate work — and many have connected those tools directly to email, file storage, and critical business systems.

Security and governance haven't kept pace. Most organisations don't know what AI is actually doing inside their environment. What tools are running, what data they're touching, and where that data is going are largely unanswered questions.

The result is an attack surface that has expanded faster than most organisations realise: tools running outside IT's visibility, data moving beyond defined boundaries, and risk accumulating quietly in the gaps.

Blocking AI isn't the answer. Governance is.

The right security controls and policy frameworks are what separate the organisations capturing the full value of AI from those unknowingly paying for it.

This playbook walks through the new AI attack surface, why traditional controls fall short, and a practical framework for building and rolling out an AI policy that works — whether you're securing one organisation or fifty.

01 • THE LANDSCAPE

AI adoption in the workplace

AI tools have shifted from optional productivity experiments to operational infrastructure — embedded in daily workflows, connected to core business systems, and used by employees across every function and seniority level.

93%

of active organisations show AI tool activity

26%

are running six or more AI tools simultaneously

59%

of employees use AI tools their organisation hasn't sanctioned

Source: Field Effect customer-base telemetry.

The pressure to adopt is real. Businesses that don't embrace AI risk falling behind, and employees feel the weight. The result is adoption that outruns security best practice: tools deployed outside IT and security oversight, without anyone realising the exposure it creates.

Employees feed unsanctioned tools client data, financial records, internal strategy, and personally identifiable information. The data risk is clear, and so are the compliance implications.

SMBs and mid-market organisations feel this most: less governance infrastructure, fewer dedicated security resources, and fewer controls to catch what's happening — all while facing the same pace of adoption as their larger counterparts.

You can't govern what you can't see. Right now, most organisations can't see enough.

02 • THE EXPOSURE

The new AI attack surface

AI doesn't just introduce new threats — it expands the terrain those threats operate across. Every tool adopted, every integration enabled, every unsanctioned application adds new surface area. Understanding where AI extends the attack surface is the first step to defending it.

01 Connected systems

AI tools integrate directly with the systems that run the business: email, file storage, CRMs, calendars, and collaboration platforms. They are often over-permissioned for the role they're assigned. Every integration that makes an AI tool more useful also pulls another system into the attack surface.

02 Shadow AI

Even where policies exist, security teams often lack visibility into what tools are running, what data they can access, or what systems they connect to. This is the shadow IT problem at a different order of magnitude. Where unsanctioned software once meant one app on one device, unsanctioned AI can aggregate sensitive data across the entire organisation, connect to critical systems, interpret that information, and share it externally.

03 Third-party and vendor connections

Every AI tool is also a supply-chain relationship. Data shared with an AI tool moves through external vendors, cloud providers, and APIs that sit entirely outside your control. The security posture of those third parties becomes part of your risk profile — assessed or not — and can introduce conflicts with compliance and regulatory requirements.

03 · WHY LEGACY CONTROLS FALL SHORT

Why traditional security controls fall short

Traditional security wasn't built for this

Traditional security runs on deterministic logic: known patterns and signatures, predictable behaviour. That model works when software behaves the same way every time. AI doesn't.

It interprets language, context, and intent dynamically, so the same prompt can produce different outcomes. Intent can be misguided or manipulated, and the AI acts on it regardless. There is no signature to match, no deviation to flag, no reliable way for legacy tools to recognise when a model is doing something it shouldn't.

**93%**

of AI agent permission requests are approved without meaningful scrutiny.

Source: Anthropic's own data.

Permission prompts appear mid-task, are written in technical language, and carry an implicit pressure to click through. Users are trying to get work done, not conduct a security review, and most have no way to assess whether a requested permission is reasonable for the tool's stated purpose.

That's a structural problem, not a user failing. When permission prompts are the primary control between an AI agent and sensitive systems, the control isn't working.

03 · WHY LEGACY CONTROLS FALL SHORT

Where the controls fall short, in practice

▲ AI agents look like both malware and normal users

AI agents download files, search data, and make web requests using the same tools and techniques malware does. Traditional security faces an impossible choice: block legitimate AI activity, drown in false positives, or exclude agents from monitoring entirely. All three leave the organisation exposed.

▲ Exclusions create the blind spot

Most teams choose exclusions to manage the noise, so malicious commands executed through an agent generate no alerts and attract no scrutiny. Threat actors have noticed. Using an AI agent as a proxy for malicious execution is a living-off-the-land technique, and it works precisely because the agent has been removed from monitoring.

▲ Agents run as the user

AI agents operate under the user's account and permissions, including administrative access. Their activity blends into normal user behaviour, because that's exactly what it's designed to look like. Tools built around behavioural deviation have no baseline to work from.

▲ No malware needs to be installed

Malicious scripts can execute entirely in memory, leaving nothing for antivirus to detect. The attack completes, and traditional endpoint protection never had anything to catch.

▲ A compromised agent doesn't wait

A compromised agent can reason, pivot, and escalate access autonomously, often completing an attack chain before an analyst can open a ticket. The speed of agentic AI eliminates the window traditional detection depends on.

INCIDENT REPORT · APRIL 2026

How a trusted AI agent became an attack vector

CASE STUDY — FIELD EFFECT MDR

In April 2026, Field Effect MDR detected and responded to an attack that illustrates exactly why traditional security controls fall short against AI-driven threats.

A developer was running Cursor, a common AI coding agent, on a macOS endpoint. Via social engineering, the threat actor convinced the developer to enter a malicious prompt — likely copied from a fake troubleshooting guide surfaced at the top of a web search result. No exploit or jailbreak needed. Agents are already permitted to download and run scripts, so that was all it took.

From that single prompt, the AI agent attempted to autonomously execute the full attack chain by:

- | | |
|--|---|
| 01 Downloading a two-stage malicious script disguised with a legitimate-looking URL | 02 Stripping file attributes to bypass macOS security controls |
| 03 Evading sandbox detection | 04 Copying credentials, session cookies, cryptocurrency wallets, SSH keys, and cloud access tokens |
| 05 Exfiltrating the data to an attacker-controlled server in under two minutes | 06 Installing a persistent backdoor with system-level permissions for long-term access |

In this example, every command was executed by the trusted AI agent, and admin prompts appeared to come from the agent itself. Traditional endpoint protection had nothing to detect.

Field Effect analysts identified the anomalous behaviour, reverse-engineered the obfuscated scripts, isolated the endpoint, and contained the threat.

Read the full report: [Field Effect detects AMOS Stealer delivered via Cursor](#) →

04 • THE FRAMEWORK

What strong AI controls look like

A strong AI policy doesn't need to be complex. It needs to be clear, enforceable, and built on the right foundations: guardrails that let you capture the benefits of AI while reducing exposure when mistakes happen, data is mishandled, or outputs can't be trusted. Five controls provide that foundation.

01 AI asset inventory

The complete picture of every AI tool in use: who owns it, what business purpose it serves, and what data it can access. Most organisations discover significantly more AI usage than expected once they begin.

THE KEY QUESTIONS

- ▲ Which AI tools are in use, sanctioned and unsanctioned?
- ▲ What business purpose does each tool serve?
- ▲ Are any tools on free tiers with different data-retention terms?
- ▲ How often is the inventory reviewed and updated?
- ▲ Who owns each tool and is accountable for its use?
- ▲ What data can each tool access, and does it need that access?
- ▲ Is there a process to request approval of new tools?
- ▲ Do any tools have MCP server integrations that extend their access?

04 • THE FRAMEWORK

02 Acceptable use policy

The rules employees need to work confidently and safely with AI: which tools are approved, what data can be shared, and what's off-limits. Without clear expectations, employees make their own judgement calls, and exposure follows.

THE KEY QUESTIONS

- ▲ Which tools are explicitly approved, and which prohibited?
- ▲ Are client data, financial records, PII, and IP defined as restricted?
- ▲ What happens when someone wants a tool not yet approved?
- ▲ What constitutes misuse, and what are the consequences?
- ▲ What data can be shared with AI tools, and what is off-limits?
- ▲ May employees use personal AI accounts for work?
- ▲ Do rules differ by role, department, or employment type?
- ▲ How will the policy be communicated and acknowledged?

03 Access and least privilege

AI tools connected to email, file storage, CRMs, and collaboration platforms should have access only to what they strictly need. Over-permissioned tools are among the most common and most preventable sources of exposure: a single compromised tool with broad access doesn't create a limited incident, it creates a gateway.

THE KEY QUESTIONS

- ▲ Does each tool need its current level of access to function?
- ▲ Are permissions reviewed regularly, and on what cadence?
- ▲ How is access revoked when a tool is decommissioned?
- ▲ How are permission changes tracked and logged?
- ▲ Who granted permissions, and were they reviewed first?
- ▲ Do any tools hold admin access that could be reduced?
- ▲ Is there a least-privilege standard tools must meet to onboard?
- ▲ Can access be revoked quickly during an incident?

04 • THE FRAMEWORK

04 Human review requirements

AI accelerates work, but accountability stays human. Define which outputs require review before they're shared or acted upon — especially customer communications, financial analysis, legal content, and security recommendations. The risk isn't that AI gets things wrong occasionally; it's that no one defined when a human needed to check.

THE KEY QUESTIONS

- ▲ Which outputs require human review before being acted upon?
- ▲ How are errors identified, corrected, and documented?
- ▲ How is accountability documented when output is acted on unreviewed?
- ▲ Who is responsible for reviewing each category?
- ▲ What is the escalation path for a significant error or harm?
- ▲ How are review requirements communicated to new staff?

05 Incident response preparedness

AI incidents — data exposure through an unsanctioned tool, unauthorised use, hallucinated output acted on, or compliance violations from data shared externally — are new scenarios that legacy response plans won't account for. Effective response needs rapid detection and containment that traditional workflows weren't designed to provide.

THE KEY QUESTIONS

- ▲ What constitutes an AI-related security incident here?
- ▲ What are the escalation paths for different incident types?
- ▲ Does your IR plan account for AI-specific scenarios?
- ▲ Have AI scenarios been in tabletop exercises?
- ▲ Who responds, and is that clearly documented?
- ▲ How fast can access be revoked from a suspect tool?
- ▲ Are AI tool activity logs maintained to support investigation?
- ▲ Are clients notified if their data is exposed, and how?

05 • IMPLEMENTATION

How to build and roll out an AI policy

The policy itself is relatively straightforward. The harder work is building the foundation it depends on: understanding your security posture, defining risk tolerance, and assigning clear ownership. Without that groundwork, even a well-written policy won't hold.

▲ Assess your security maturity first

AI amplifies weaknesses that already exist. A small permissions gap that went unnoticed for years can be blown wide open the moment an AI tool starts aggregating data across systems it technically has access to. The more mature the program, the faster and more confidently you can move.

▲ Start with risk tolerance, not tools

Guardrails reflect the organisation's risk tolerance. Before writing policy, leadership weighs the risks of AI against its benefits and decides how much risk to accept; the security team then works within that tolerance. No formal risk review process? AI adoption is reason enough to start one.

▲ Keep your policy suite lean

Resist a new policy for every consideration. Acceptable use, change management, and supplier onboarding already have a home in existing documentation. Update what exists rather than starting from scratch. Fewer policies stay fresher, get used more, and are easier to enforce.

05 • IMPLEMENTATION

▲ Vet every vendor before you connect them

Every AI-powered tool is a supplier relationship. Before onboarding, run a formal vetting process. Clients increasingly expect this governance to flow down through the supply chain, making how you govern your own AI a factor in whether you win and keep business.

THE KEY QUESTIONS

- ▲ Where is data stored, and does storage conflict with regulatory requirements?
- ▲ Do they undergo third-party auditing?
- ▲ Does the vendor maintain a security and compliance program?
- ▲ What are the data-retention and deletion terms?

▲ Define who owns AI risk

There's a distinction between who sets the risk appetite and who manages and enforces it day to day. Both roles need owners, and both should be named in the policy. For MSPs, helping clients assign accountability and build oversight capacity is a direct service opportunity.

06 • ADOPTION

Making your AI policy stick

Building the policy is the starting point. Getting it adopted, followed, and maintained over time is where most organisations fall short. The difference comes down to a few fundamentals.

Leadership must model the behaviour

Adoption fails when the policy reads as an IT initiative rather than an organisational priority. When leadership complies visibly and champions it, employees follow. If the expectation doesn't apply at the top, it won't stick anywhere.

Involve employees in the process

Policies developed with input from the people who use them earn higher adoption. Bringing representatives from key business units into development increases buy-in and surfaces practical gaps before the policy goes live.

Review and update regularly

The AI landscape moves too fast for a static policy. Set a defined review cadence — annually at minimum, more often if possible — so the policy keeps pace with new tools, threats, and regulatory requirements.

Explain the reasoning, not just the rules

Employees who understand why guardrails exist are far more likely to follow them. Fold guidance into existing security-awareness materials and frame it around real consequences. Compliance comes from understanding, not memorisation.

Enforce from day one

A policy without enforcement is a paper artifact. Configure monitoring before rollout, not after the first incident, so you know which tools are in use and whether unsanctioned activity is occurring. A useful test: if an employee violated this policy today, would you know?

Make reporting easy and safe

Give employees a clear, low-friction way to report AI concerns, and a culture where doing so is encouraged, not penalised. A no-blame reporting culture surfaces issues before they escalate.

CYBERCADDY AIDR

How CyberCaddy can help

Effective AI governance starts with a reliable security foundation. Without proper visibility, any policy operates blind. CyberCaddy's managed detection and response provides that foundation — with intelligence-grade detection, monitoring, and response that give organisations the confidence to adopt AI.

CyberCaddy AI Detection and Response (AIDR) extends the coverage you already have with CyberCaddy to the AI layer — no new agents, no separate dashboards. It gives you visibility into AI usage, enforces control, and supports stronger AI governance across your environment.

DISCOVER

Uncover shadow AI

Where AI is being used, and who's using it — including unsanctioned or unknown applications.

GOVERN

Enforce policy & control

Establish and enforce guardrails to reduce risk, eliminate shadow AI, and take back control.

ADOPT SAFELY

Support secure adoption

Monitor adoption of new AI tools so your team can support secure, responsible use across the organisation.

"CyberCaddy provides clear risk visibility, practical controls, and guidance that is technically sound and realistic — supporting informed decision-making and long-term cyber resilience rather than creating dependency."

Ben Knight — Business Systems Manager, Pinnacle Safety and Training

IN CLOSING

Embrace AI, but securely

AI is an opportunity, and the organisations that capture it most effectively will have built the foundation to do so safely. Adoption has already happened. The work now is catching governance up to the reality of how AI is being used.

A clear AI policy removes the uncertainty that makes adoption risky. Employees know what's approved. Security teams know what's running. Leadership has confidence that productivity gains aren't coming at the cost of data, compliance, or trust.

See AIDR in your environment

Premium protection shouldn't require a bigger team. Stop chasing alerts and start seeing outcomes — get a guided walkthrough of AI Detection and Response.

[Book a demo →](#)

ABOUT CYBERCADDY

We started CyberCaddy to give IT and business leaders an independent, straight-shooting cyber partner to turn to — not another IT generalist or software reseller.

Prior to starting CyberCaddy, we saw all of the issues the technology industry has: vendor kickbacks, generalist cyber skillsets, "one shoe fits all" cyber packages. Frameworks applied because they were familiar, not because they fit.

CyberCaddy exists to do it differently. We're invested in your tangible cyber security outcomes, not locked to any one vendor, framework or solution. We measure success by how much more capable our clients are to prevent and respond to cyber incidents, not by how long we can extend contracts.

WEB www.cybercaddy.com.au
EMAIL enquiries@cybercaddy.com.au

