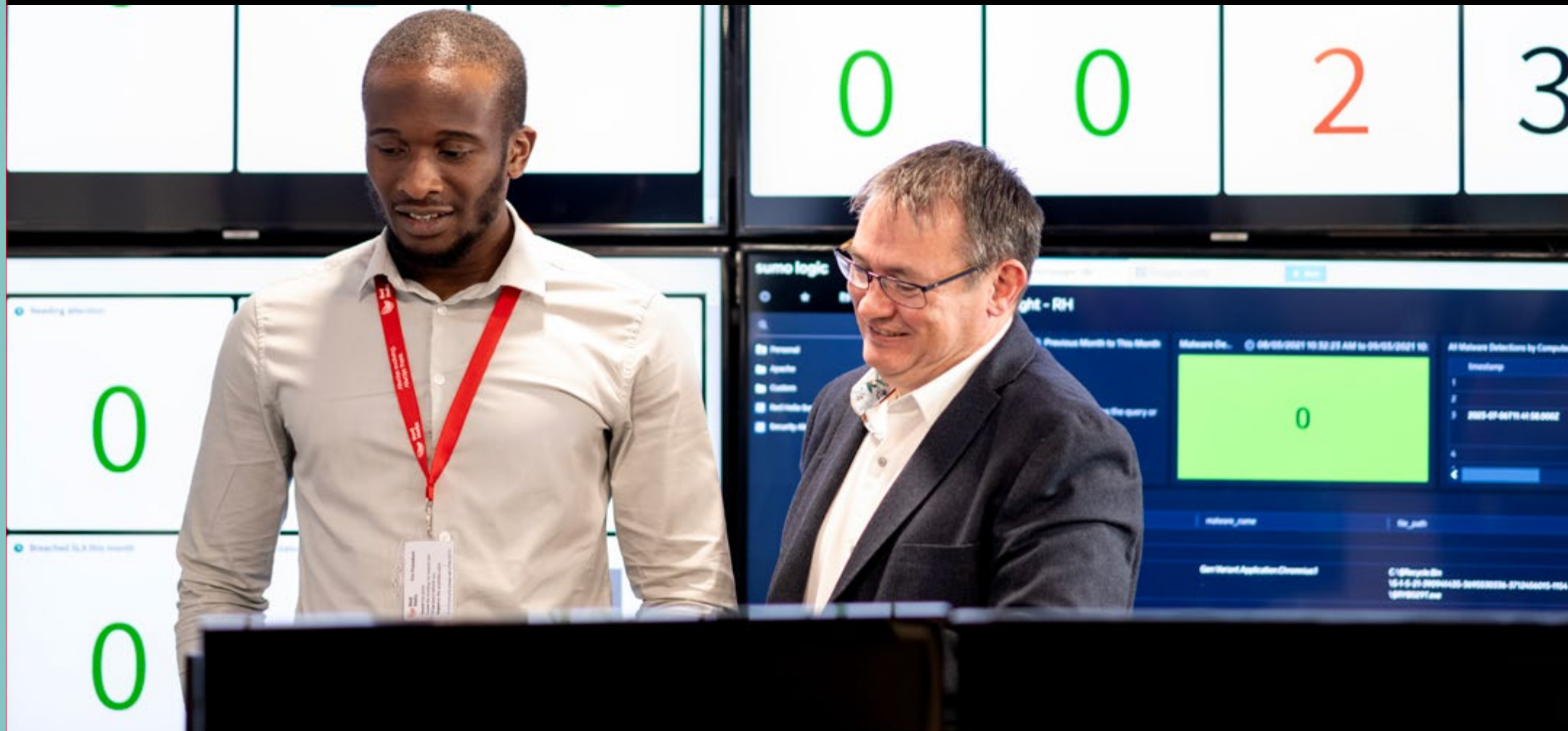
The AI Security Gap: How to find and control shadow AI before it becomes an issue



CONTENTS

The rise of Shadow AI and invisible adoption 3

Shadow AI 4

The operational risks created by unmanaged AI usage 5

Governance Versus Operational Reality 7

Combined and layered approach to AI control 8

Evaluate: understanding exposure

Execute: operationalising control frameworks

Evolve: continuous oversight and adaptation

Strategic implications for leadership 9

The Rise of Shadow AI and Invisible Adoption

AI has become one of the fastest adopted enterprise technologies in modern business history. Across every sector, employees are using generative AI to analyse information, draft communications, automate routine tasks, and support decision-making. Business leaders are increasingly driving its adoption to build competitive advantage, reshape operations and create efficiencies.

According to recent research, 88% of organisations now report using AI in at least one business function, up from 78% the previous year. What began as experimentation has rapidly become embedded in day-to-day business activity.

Historically, enterprise technologies were introduced through structured programmes involving procurement, governance reviews, security assessments, architecture approval, and controlled deployment. AI has largely bypassed this process. Employees have adopted AI tools independently, departments have embedded them into workflows, and AI agents are beginning to interact directly with corporate systems, often before governance and security teams have established visibility.

The challenge facing leadership teams is understanding where AI is being used, what risks it introduces, and how governance can evolve from policy documents into operational oversight.

Shadow AI

Shadow AI is the term being used to describe AI tools, services and usage within an organisation that has not been formally approved or integrated into risk management considerations.

Unlike malicious insider activity, Shadow AI is rarely driven by bad intent. Employees adopt AI because it helps them work faster. Marketing teams use it to create content. Finance teams use it to analyse information. Developers use it to generate code. Project teams use it to draft documentation. The attraction is straightforward: AI can reduce effort, accelerate tasks, and remove friction from routine work.

The challenge is that convenience is rarely associated with governance. Recent research highlights the scale of this behaviour. In the UK, 71% of employees report using unapproved consumer AI tools in the workplace, with more than half doing so on a weekly basis.

At the same time, 32% said they were concerned about the privacy of company or customer data inputted into consumer AI tools, and only 29% were concerned about the security of their organisation's IT systems.

This indicates a material disconnect between AI usage and the associated security risks.

For leadership teams, this creates a structural governance challenge. AI is being integrated into business workflows through individual behaviour rather than organisational strategy, which limits visibility, consistency, and control over how data is processed and where it is exposed.

The strategic question is no longer centred on adoption timelines. It is whether governance, security, and operational oversight can be re-established in an environment where AI usage is already widespread, decentralised, and largely invisible to traditional control frameworks.

The decision over whether AI should be permitted has already been made by the workforce. The pressing questions now is whether security and governance can catch up with adoption before AI is embedded in business critical processes.

"Most organisations believe they are at the beginning of their AI journey. In our experience many organisations find employees have been using AI for months or years before governance teams gain visibility of the scale. The first challenge is discovering how much AI adoption has really taken place."

– Taras Sachok, Senior Risk Consultant, Risk Crew

The Operational Risks Created by Unmanaged AI Usage

For decades, enterprise security strategies have been built around a relatively simple assumption: threats originate outside the organisation. Firewalls, web gateways, email security, endpoint protection, and network monitoring were all developed to prevent malicious actors from gaining access to systems and data.

Today, the risk is not necessarily an external attacker attempting to breach the perimeter. Increasingly, the interaction itself becomes the risk. Employees are voluntarily entering information into AI platforms. AI agents are being granted access to business systems. Decisions are being influenced by models that operate beyond the visibility of traditional security tools.

This creates an entirely new operational layer that most organisations security stack are not equipped to monitor.

Governance policies can define acceptable use, but they cannot observe whether those policies are being followed. As AI adoption accelerates, organisations are discovering that many of their existing controls stop precisely where AI interactions begin. The result is an increased risk of a data breach and increasing security exposure.

The operational exposures created by employee AI usage include:

Sensitive information entered into external AI tools
Employees may paste confidential material such as customer records, contracts, or financial data into public AI platforms to speed up tasks. Once submitted, that data may be processed outside organisational boundaries, this creates blind spots in understanding what information is leaving the organisation and how it is being processed. For organisations operating within regulated sectors, this can create exposure under GDPR, contractual confidentiality obligations, industry-specific regulations and intellectual property protections.

AI agents interacting with internal systems without full traceability
The emergence of agentic AI represents a significant shift from information generation to action execution. Unlike traditional generative AI tools that produce outputs for human review, agents can perform tasks directly within business systems, tasks such as querying databases or triggering workflows.

Without detailed traceability, it becomes difficult to determine what prompted an action, what data was accessed, or whether the outcome aligns with intended controls. Tools such as Claude Code, Claude Cowork, and Open Claw can execute shell commands, manage files, rename directory structures, write code and automate workflows all without interacting with any security controls.

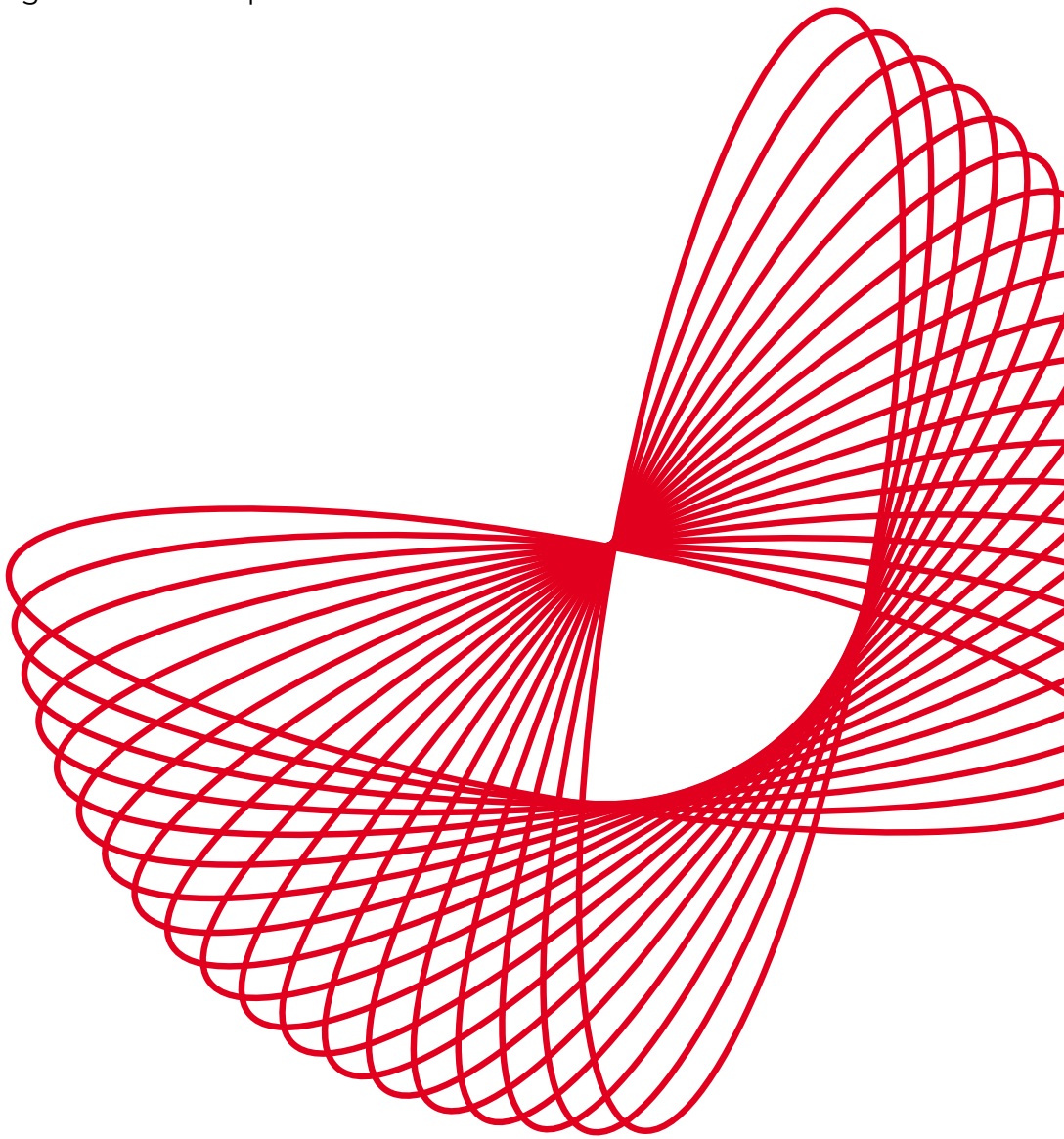
Lack of visibility into where AI is being used across the organisation
AI adoption frequently occurs at team or individual level without central registration or approval. This leads to fragmented usage patterns across departments, making it difficult to build a reliable inventory of AI touchpoints. The absence of visibility limits the ability to assess risk exposure or apply consistent governance standards.

Exposure to prompt injection techniques that manipulate model behaviour
Prompt injection occurs when malicious or unintended instructions are embedded within inputs or connected data sources, influencing model outputs. These techniques can cause AI systems to reveal restricted information or perform actions outside intended boundaries. The risk is heightened where AI is connected to internal datasets or automated workflows.

Governance policies that cannot be consistently enforced at point of use
Even where AI policies exist, they are often applied through guidance rather than technical enforcement. Users may not encounter control checks when they interact with AI tools, so it is up to personal discretion, leading to inconsistent adherence.

Limited capability to investigate AI-related incidents once they occur
When AI systems are involved in data exposure or operational anomalies, the absence of detailed interaction logs makes forensic investigation difficult. Traditional incident response tools may not capture prompt histories, model outputs, or agent decision paths. This limits the ability to determine root cause or implement corrective actions.

Fragmented accountability for AI use across teams and functions
Responsibility for AI usage is often distributed across IT, security, data, and individual business units without clear ownership boundaries. This fragmentation can result in gaps where no single function is accountable for oversight or risk management. It also complicates escalation paths when issues arise, slowing organisational response.



Governance Versus Operational Reality

Many organisations have developed AI policies, ethical frameworks, and usage guidelines. These provide structure at a documentation level but do not necessarily translate into operational control.

Despite governance and guidelines being in place, many employees will be using AI anyway. Even if you block all LLM's on the network it does not stop employees finding loopholes like emailing themselves information running it through AI on a personal device and sending the output back to themselves.

Traditional governance approaches assume that there are enforceable controls. AI usage disrupts this assumption by leaving the controls in the hands of the employee. Security tooling designed for endpoint, network, and application layers does not naturally extend to prompt-based interactions or model-mediated workflows. This creates a visibility gap between policy intent and real-world execution.

AI operational resilience therefore depends on knowing where AI is being used, continually monitoring AI interactions and having mechanisms in place to detect and respond to AI-related anomalies or misuse.

Without these elements, governance remains declarative rather than operational and impossible to monitor or measure the true risk exposure.

Regulation Is Catching Up

While organisations continue to adapt to AI adoption, regulators are moving quickly to establish expectations around governance and accountability.

Frameworks such as ISO 42001, the EU AI Act, UK Government AI governance principles, Information Commissioner's Office guidance, and National Cyber Security Centre recommendations all point towards a common direction of travel: organisations will increasingly be expected to demonstrate that they understand how AI is being used, how risks are assessed, how decisions are governed, and how controls are monitored.

For many organisations, ISO 42001 is emerging as a practical framework for establishing these foundations. The standard provides a structured approach to AI governance, risk management, accountability, and continual improvement. Increasingly, organisations are viewing it not simply as a certification exercise, but as a way to bring structure and consistency to AI oversight.

However, certification alone does not provide continuous operational visibility.

Governance frameworks can define what should happen. Organisations still require mechanisms to understand what is actually happening.

Building a layered approach to AI control

Red Helix helps organisations secure AI adoption through our ethos of Evaluate, Execute, Evolve. This model reflects the progression from discovery to operational integration and ongoing oversight.

Evaluate: understanding exposure

The initial phase focuses on establishing visibility across the organisation's AI usage landscape and exploring the regulatory setting for the business and its implications.

Typical activities in this phase include:

- AI governance and risk workshops
- Assessment of AI usage across business units
- Review of data exposure pathways through AI tools
- Security testing focused on AI interaction points

The output is a structured view of exposure and governance gaps, prioritised by operational and regulatory impact. This also means your organisation will be in a good position to consider getting ISO 42001 certified.

Execute: operationalising control frameworks

Policy development alone does not translate into control over AI usage. This phase focuses on embedding governance into operational systems and technologies.

AI Detection & Response (AIDR) is designed to monitor, detect and respond to threats involving AI systems, AI interactions and generative AI usage. This includes prompt injection attacks, sensitive data exposure, shadow AI usage, unsafe AI interactions, malicious AI agent behaviour, unauthorised AI tools and AI-related operational risk.

CrowdStrike Falcon AIDR provides visibility of who is using AI and how. It then further extends visibility into prompts, AI agents, workflows and AI usage patterns, helping organisations identify and block threats and suspicious behaviour before they become incidents.

As well as implementing AIDR, AI awareness training is needed to ensure employees are clear on the risks of using AI at work and home.

The focus shifts from defining rules to embedding them into operational behaviour and system design.

See AIDR in action here.

Evolve: continuous oversight and adaptation

AI usage changes rapidly as new tools, models, and agent capabilities are introduced. Governance frameworks require continuous adjustment to remain aligned with operational reality.

So, continually having ongoing oversight and reviews is essential for staying vigilant.

This phase reflects the transition from static governance to adaptive operational oversight.

"Traditional security monitoring was built around users, devices and applications. AI introduces a new layer that requires dedicated monitoring and response capabilities."

Tom Exelby, Head of Cyber, Red Helix

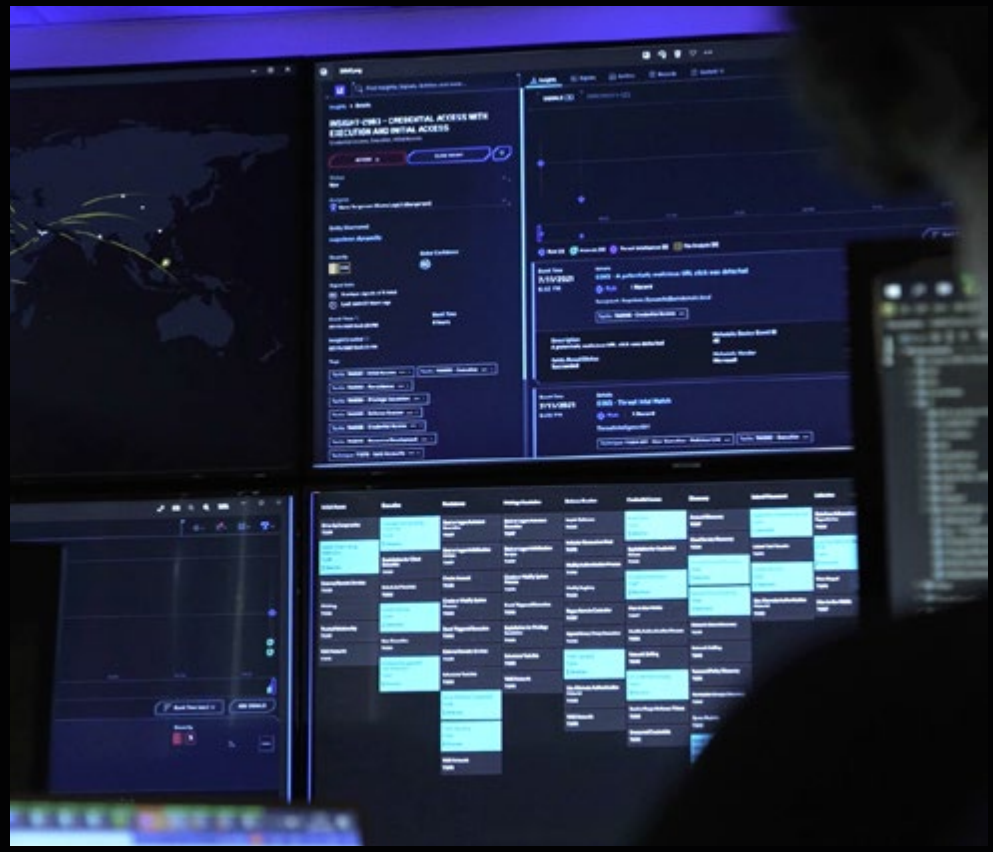
Strategic implications for leadership

The introduction of AI across organisations means that monitoring security risk has evolved beyond application boundaries and extended into user interaction layers that are too granular for traditional security monitoring tools.

From a leadership perspective, this introduces several considerations:

- Existing security models may not provide coverage for AI-mediated data flows
- Policy frameworks may not translate into enforceable operational controls
- Incident response capabilities may lack visibility into AI-driven activity

The organisations that manage AI risk most successfully are unlikely to be those that restrict adoption. They will be those that establish visibility, accountability and operational oversight while enabling AI to be used safely across the business.



Achieving AI control

AI adoption is already happening inside most organisations. The question is whether IT and security teams have the visibility, control and evidence needed to support it safely.

Red Helix helps organisations move from AI policy to operational oversight through AI risk assessment, AI Detection and Response, security monitoring and ongoing control improvement.

Book a short demo to see how you can identify Shadow AI, detect risky usage, monitor AI interactions and build a clearer picture of AI risk across your organisation.

Contact Red Helix today and revolutionise your IT security.

+44 (0)1296 397711

info@redhelix.co.uk

Phoenix House
Smeaton Close
Aylesbury
Buckinghamshire
HP19 8UW

Red Helix

redhelix.com