MERIT

The Shadow Agent Problem: What Happens When Every Employee Starts Deploying AI Agents on Your Enterprise Systems Without Telling You

The next wave of enterprise AI risk is not the agents your IT team built. It is the ones your finance, sales, and operations teams have already deployed without telling anyone.

Authored by Tharun Mathew
Head of Data & AI Solutions, Merit Data and Technology

CONTENTS

Table of Contents

EXECUTIVE SUMMARY

SITUATION

Shadow IT was about employees signing up for SaaS tools without telling anyone. Shadow AI was about employees pasting customer data into ChatGPT. Shadow agents are something else again. They are autonomous systems that an employee in finance, sales, or operations has wired up over a weekend, given access to company data, and pointed at production workflows. They take actions. They send emails. They update records. They write to systems. And they do it without a security review, without a procurement record, and almost always without anyone in IT or risk knowing they exist.

COMPLICATION

The tooling has made this trivial. Cursor, Windsurf, ChatGPT Custom GPTs, Claude with MCP connectors, n8n, Make, Zapier with native AI, GitHub Copilot, OpenAI Assistants. A motivated analyst can build an autonomous workflow in an afternoon that connects to the CRM, pulls customer data, drafts and sends communications, and updates records. Twenty years of building enterprise data infrastructure tells me the controls were not designed for this. Traditional security tools suffer from protocol blindness: they can see that a model API was called, but they cannot inspect the semantic intent of what the model told the tool to do, what data it passed, or whether the action that followed was within the scope of what any human approved. A firewall rule cannot evaluate whether a tool invocation should have happened. An identity system cannot distinguish the employee from the agent acting as the employee. The governance gap is not a policy failure. It is a visibility failure, and writing more policy does not fix a problem the organisation cannot yet see.

RESOLUTION

The organisations that will get through 2027 without a major shadow agent incident are not the ones banning the tools. Bans drive usage further underground and most employees would keep using personal accounts anyway. The ones who get this right are doing three things in parallel. They are building visibility into what is actually deployed across the environment. They are giving employees sanctioned alternatives that are good enough to compete with the unsanctioned ones. And they are putting governance into the data layer rather than into policy documents. None of these are exotic. All of them require deliberate work that most enterprises have not yet started.

KEY FINDINGS

1. **Shadow agents are not shadow IT.** Shadow IT was a passive risk: an employee using an unapproved tool. Shadow agents are an active risk: an autonomous system taking actions across enterprise systems on the employee's behalf. The harm mechanism is categorically different. The governance frameworks built for the previous problem will not address this one.

2. **The visibility gap is the governance gap.** 82% of executives feel confident their policies protect the organisation from unauthorised agent actions. Only 14% have full security approval for the AI agents currently deployed in their environment. That 68-point gap is the actual governance posture, regardless of what the policy document says.

3. **Bans do not work.** Roughly half of employees would continue using personal AI accounts even after an organisational ban. Prohibition drives shadow agents deeper underground rather than eliminating them. The CISOs who are succeeding are competing for employee adoption, not legislating against it.

4. **Regulation is not waiting.** The EU AI Act's high-risk obligations take full effect in August 2026. Shadow agents operating in HR, credit, employment, or essential services fall squarely within scope, whether or not the organisation officially knows they are runnin Regulators will not accept "we did not know" as a defence, particularly when the audit trail does not exist to reconstruct what an agent did.

5. **The fix is data infrastructure, not policy infrastructure. not a compliance exercise.** Most enterprise AI governance programmes are policy documents and committees. Shadow agents are a runtime problem. Closing the gap requires investments in agent discovery, identity management, and behaviour monitoring at the data layer. Policy without runtime enforcement is performative.

KEY STATISTICS

80%

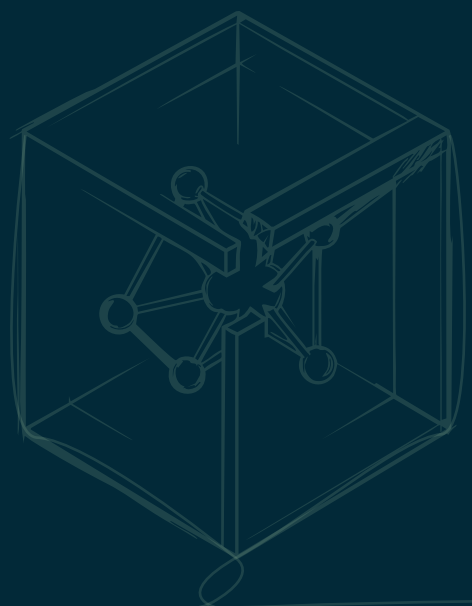
of organisations have already encountered risky agent behaviours, including unauthorised data exposure (McKinsey, 2025)

1 in 5

companies has a mature governance model for autonomous AI agents (Deloitte, 2026)

40%+

of agentic AI projects will be cancelled by end of 2027 due to inadequate risk controls (Gartner)



Section 1

Why Shadow Agents Are Not Shadow IT

The pattern is consistent across every enterprise I have spoken with in the last year. A Custom GPT wired to the CRM, running quietly in the marketing function. Cursor with read access to the entire codebase, including the parts with embedded credentials, running in finance. Zapier triggering outbound emails based on conditions pulled from three separate SaaS systems, running in operations. No security review. No procurement record. No IT visibility. By the time it surfaces, it has been running for months and is genuinely useful, which is precisely what makes it hard to address.

This pattern has a name now: shadow agents. It is the natural evolution of shadow IT and shadow AI, but the harm profile is categorically different from either. Shadow AI was primarily a read risk: data leaving the organisation through a prompt. Shadow agents introduce state change persistence: autonomous systems modifying production records, sending communications, updating databases, and triggering downstream workflows, all without an audit trail. When something goes wrong, the cost is not just the exposure. It is the reversal: finding everything the agent touched, reconstructing what it changed, and unwinding actions that other systems have already acted on. The governance frameworks built for the previous generation of risk were never designed for that failure mode.

The Three Generations of Unsanctioned Technology

Shadow IT is the well-understood version. An employee signs up for Dropbox to share files. An engineer uses a personal GitHub for a side project that quietly becomes a production dependency. A team adopts Slack before security has approved it. The risk profile is reasonably bounded: data ends up in places it should not be, access controls are inconsistent, and there is no audit trail. Bad. Manageable, with the right CASB tooling and a procurement process that bites.

Shadow AI is the next generation. An employee pastes a customer email into ChatGPT to get a faster reply. A developer drops production code into a personal Claude account to debug it. A finance analyst asks an LLM to summarise a board pack. The data leaves the organisation and ends up training a model owned by someone else, or sitting in a vendor's logs, or visible to whoever has the right URL. The harm is real, the regulatory exposure is bigger than shadow IT, and the existing tools mostly cannot see it because the data leaves through prompts rather than file uploads. Still, in most cases, the harm is bounded by what the employee chose to share.

Shadow agents are not bounded that way. An autonomous system, once it has access to your CRM, your email, your calendar, your codebase, your finance system, will act on whatever it finds there. It will send emails on the employee's behalf. It will update records. It will trigger workflows in connected systems. It will keep doing this whether or not the employee is paying attention. Google Cloud's 2026 Cybersecurity Forecast describes shadow agents as creating "invisible, uncontrolled pipelines for sensitive data, potentially leading to data leaks, compliance violations and IP theft." That framing is right. The pipelines are the issue. The agent is the operator at the end of them.[1]

"Agency isn't a feature. It's a transfer of decision rights. The question shifts from "is the model accurate?" to "who is accountable when the system acts?""

McKinsey, State of AI Trust 2026

The Harm Mechanism Is Different

With shadow IT, the worst case is data exfiltration. Bad, but mostly recoverable through detection, containment, and the standard incident response playbook. With shadow AI, the worst case is data exposure to an external model, which is worse but still primarily a data security problem. With shadow agents, the worst case is autonomous action across multiple production systems, taken by something the security team did not know existed, against data the agent should not have had access to, with consequences the organisation is going to have to unwind manually because the agent did not log what it did.

This is why every major analyst house has shifted its framing in the last twelve months. McKinsey's research is specific: 80% of organisations have already encountered risky behaviour from AI agents, including unauthorised data exposure and improper system access. Gartner predicts that more than 40% of agentic AI projects will be cancelled by the end of 2027, and that AI-related legal claims will exceed 2,000 by year-end 2026, driven in significant part by inadequate risk guardrails. The Cloud Security Alliance's AI Controls Matrix added 18 security domains and over 240 control objectives specifically because the existing controls did not cover what agents do.[2][3][4]

Why Existing Tooling Cannot See This

- **CASB and SaaS management tools were built for SaaS apps.** An agent built in Cursor, running in a developer's local environment, calling out to api.anthropic.com or api.openai.com over a personal API key, is not a SaaS app. It does not show up in the procurement record because nobody bought anything. The OAuth scopes it requests look like the scopes any legitimate developer tool would request. The only artefact in the network logs is calls to a model API endpoint, which the company itself probably uses for legitimate purposes elsewhere.
- **DLP was built to inspect data.** Data loss prevention tools watch for patterns: credit card numbers, social security numbers, certain keywords in outbound communications. They do not inspect what an agent is doing. They cannot evaluate whether a tool invocation should have happened. They cannot tell you whether an agent that just sent an email was authorised to send it on whose behalf.
- **Identity systems were built for human users and known service accounts.** Most enterprise identity infrastructure assumes the actor is a person or a defined non-human service. An autonomous agent built by an employee, authenticating as that employee, calling APIs on their behalf, fits neither category cleanly. The agent has the employee's full access. There is no audit trail that distinguishes "the employee did this" from "an agent did this on the employee's behalf". For a regulator asking how a particular customer email got sent, that distinction matters.
- **AI governance committees were built for sanctioned AI.** The committee approves the model. The procurement team licenses the platform. The security team reviews the integration. None of this applies to an agent that nobody asked anyone to approve. The committee is solving a different problem to the one that is actually creating exposure.

The Compounding Problem with Doing Nothing

Every quarter that goes by without a shadow agent governance programme, two things happen. First, more agents get deployed. The tooling is getting easier, not harder. Cursor's agent mode is a default now. ChatGPT Custom GPTs and Claude Projects with MCP connectors are weekend projects. Zapier and n8n have native AI integrations. The trajectory is more shadow agents, not fewer.

Second, the cost of getting visibility goes up. An agent that has been running for six months has dependencies. Other people use its outputs. Other systems have come to expect what it does. Pulling it out is not a question of revoking a credential. It is a question of finding everything downstream that has come to rely on it and figuring out what to replace. The retrofit cost compounds. The organisations that started this work in 2025 are in a different place to the ones starting it in 2026, and the gap will widen further by 2027 when the EU AI Act's high-risk obligations are in full force.

Section 2

How Employees Are Building Agents Faster Than IT Can See Them

To understand the scale of what is happening, it helps to look at what the average motivated employee in 2026 actually has access to. The tools have moved from "emerging category" to "weekend project" in eighteen months, and most CISOs I speak with are still calibrating their threat models against where the market was a year ago.

The Tooling Has Become Trivial

Cursor and Windsurf are AI-powered IDEs that include agent modes. A developer using either of these is not just getting code completion. They are deploying an autonomous system with read and write access to their development environment, authenticated as them, capable of generating code, executing it, modifying configuration, and interacting with version control. The agent operates with the developer's full permissions. It does not show up as an agent in any inventory.

ChatGPT Custom GPTs and OpenAI's Assistants API let users create AI agents that call external APIs. Employees are building Custom GPTs that access Salesforce, pull from internal databases via custom connectors, and write back results, all configured in a consumer-grade UI without IT involvement. Anthropic's Claude with MCP connectors is the more flexible version of the same pattern: an employee can wire up Claude to read from Gmail, write to Asana, query a Postgres database, and execute commands on a remote server, often using credentials they already have for their day job.

Workflow tools have caught up too. n8n, Make, and Zapier all offer native AI agent capabilities. What used to require a small engineering team now takes an afternoon. A finance analyst with no formal coding background can build an autonomous workflow that monitors a shared inbox, classifies incoming invoices, extracts the line items, validates them against a database, and pushes the results into the accounting system. None of this requires IT involvement. The employee is solving a real problem, often a problem IT could not solve fast enough through formal channels.[5]

> **"Modern agentic architectures don't just run a single autonomous workflow. They chain them. An employee might deploy a research agent that triggers a drafting agent that triggers a send agent. Each step has its own permissions footprint. Each handoff is a potential data exposure."**
>
> Industry analysis, 2026

Why Employees Build Them

The temptation is to frame shadow agents as a security failure or an employee compliance failure. That framing misses the point. Employees build agents because the tools have become accessible enough to use, and because the alternative is waiting for IT to deliver something equivalent that may take six months and may not come at all. Productivity pressure is real. The agents work. The work gets done faster. The employee gets credit. The risk lives somewhere else, on someone else's balance sheet. And critically, the credentials needed to make those agents useful are already in the developer's environment: corporate session tokens sitting in local configuration files, inadvertently passed into Cursor, Windsurf, or similar tools, giving the agent production-level access nobody explicitly approved.

Industry research has been clear about the motivations. Speed and productivity are the top reasons. Inadequate sanctioned alternatives is second: when enterprises do not provide AI tools that match what employees can build for themselves, employees build them. Absent or unclear policies is third: when the organisation has not told people what is acceptable, they make their own decisions. None of these are character flaws. They are predictable responses to a tooling environment that has changed faster than the governance environment around it.

82%

of executives feel confident in their AI policies, but only 14% have full security approval for deployed agents

60%

of organisations have experienced data exposure events linked to employee use of public AI tools

47%

of employees access AI through personal accounts, completely outside enterprise controls

The Discovery Problem

Once you accept that shadow agents are being built, the operational question is whether you can find them. The honest answer for most organisations is no, not yet, and the longer it takes to acquire that capability the harder the problem gets. Discovery is not a tool. It is a programme. The signals are spread across multiple layers of the stack and no single tool will surface all of them.

Network traffic to model API endpoints is one signal. Calls to api.openai.com, api.anthropic.com, and similar destinations from devices that should not be making them, at volumes that suggest something more than casual use, are detectable. OAuth tokens issued to applications that are not on the approved software list are another. Service account behaviour that does not match its declared purpose is a third. Personal AI subscriptions expensed through corporate cards is a surprisingly common fourth. None of these are foolproof. All of them together start to give you a picture.

The fastest path to discovery, in most engagements I have seen, is asking. A structured AI tool disclosure process, framed as "help us understand what you are using so we can support you better, not penalise you," surfaces the majority of shadow agent usage quickly. Employees mostly do not want to be hiding things. They want to keep the productivity benefit. If the organisation makes it cheaper to disclose than to hide, most of the agents come into the light voluntarily. That cannot be the whole programme, but it is a reasonable first step.[7]

Section 3

The Governance Gap and Why Policy Will Not Close It

Most enterprises I have looked at in the last twelve months have an AI governance policy. Many have an AI committee. Some have published acceptable use guidelines. None of these things, on their own, have closed the shadow agent gap. The reason is structural. Policy operates on intent. Shadow agents operate at runtime. The two layers are not connected in most enterprises, and writing more policy does not connect them

The 68-Point Gap

Industry research from 2026 has put a number on what most CISOs already feel intuitively. Roughly 82% of executives feel confident their policies protect against unauthorised AI agent actions. Only around 14% of those same organisations have full security approval for all the agents currently deployed in their environment. The gap between perceived governance and actual governance is roughly 68 percentage points, and it is the gap most enterprise risk programmes are not yet measuring.[6][8]

This is not unique to AI. It is the same gap that exists in any domain where policy ambition outruns operational measurement. The difference with shadow agents is that the actions happen at machine speed, leaving very little time for human intervention to compensate for the missing controls. Policy that depends on humans noticing and reporting will not work because the humans cannot see what is happening fast enough. Governance has to be embedded into the runtime layer, where the agent is actually doing things.

> "Most organisations have policies, committees, and training in place, but lack mechanisms that operate in real time at the point where AI risk is actually created: prompts, uploads, and embedded AI features inside SaaS tools."
>
> CultureAI / Censuswide research, 2026

Why Traditional Governance Tools Cannot Bridge It

Three structural mismatches make this hard.
First, traditional governance assumes a human in the loop. The user requests access, an approver reviews, a policy is applied, the action is logged. The whole model assumes deliberate, sequential, reviewable activity that happens at human speed. Agents collapse that timeline to milliseconds. By the time a human governance step would fire, the agent has already done what it was going to do.

Second, traditional governance assumes the actor is identifiable and singular, and this is where identity inheritance becomes the primary flaw. A user is granted permissions based on the assumption that a human will exercise them deliberately, one action at a time, at human speed. An employee-deployed agent inherits those same permissions but operates at machine speed, across multiple systems simultaneously, executing hundreds of actions in the time it would take the employee to perform one. The access was never designed to be used that way. Multi-agent systems compound this further: when one agent calls another, the inherited permissions stack in ways no access review ever modelled. Most identity systems have no clean answer for whose audit trail this is.

Third, traditional governance assumes the action set is bounded. A policy might say "this user can read but not write to system X." Agents take compound actions across systems, and the legal or operational implications of those compound actions are not always predictable from the individual permissions involved. An agent that has read access to email and write access to a calendar can, in combination, do things neither permission alone would suggest, like extract sensitive content from emails and embed it in calendar invites visible to a wider audience.

The Regulatory Pressure Is Tightening

The EU AI Act's high-risk obligations take full effect in August 2026. For any organisation operating in or selling into European markets, this is not future-state planning. It is a current compliance obligation, and shadow agents do not get an exemption because they were not formally deployed. Agents operating in HR, credit, employment, essential services, or any other enumerated high-risk domain fall within scope, whether or not the organisation officially knows they are running.[11][12]

The penalties are not theoretical. Infringements related to prohibited AI systems can lead to fines of up to EUR 35 million or 7% of global annual turnover. Lesser infringements can attract fines up to EUR 15 million or 3%. Beyond fines, regulators can mandate the withdrawal of non-compliant systems from the market, which for an organisation that has built workflows on top of agents it now has to disable, is a different kind of operational problem.[12]

The harder question is evidentiary. When a regulator or a court asks how a particular automated decision was made, the answer needs to be retrievable. Most shadow agents do not log their actions in any way that survives an audit. The agent ran. It did things. There is no record beyond what its outputs eventually wrote into other systems. Reconstructing what the agent knew when it acted, what context it had, what alternatives it considered, and why it took the action it took is impossible without infrastructure that almost no organisation has yet built. The Cloud Security Alliance's 2026 research note on AI agent governance is direct about this: the absence of evidence-quality audit trails is both a security problem and a compliance liability.[4]

Why Bans Backfire

The instinctive response from some risk leaders has been to ban the tools. This rarely works for long. Industry research shows that nearly half of employees would continue using personal AI accounts after an organisational ban. Samsung famously banned ChatGPT in 2023 after engineers leaked source code into it, then reversed the decision and built an internal alternative because the productivity gap was too large to sustain. Bans drive usage further into the shadows where it is harder to see and harder to govern. The CISOs who are getting traction on this are not the ones legislating against employee behaviour. They are the ones competing for it.[13]

Section 4

Five Pillars of a Shadow Agent Governance Programme

A shadow agent governance programme is not a single tool, a single policy, or a single committee. It is five interdependent capabilities that have to be designed to work as a coherent whole. They are not phases. They are not a maturity ladder. They are simultaneously active operational concerns, and the dependencies between them are why partial programmes tend to produce the illusion of governance rather than the substance of it. Build identity controls without behavioural monitoring and the agent operates correctly until it does not. Build policy without runtime enforcement and the policy is performative. The pillars constrain each other, which is why they have to be built together.

Pillar 1 - Discovery and Inventory

You cannot govern what you cannot see. Discovery is the foundation of every other pillar, and it is the one most organisations have invested least in. The default is to assume that procurement records and software asset management cover the agent estate. They do not, because shadow agents do not pass through procurement and most do not show up as software.

A serious discovery programme operates at multiple layers simultaneously. Network telemetry surfaces calls to model API endpoints, MCP servers, and agent framework infrastructure. Identity logs surface OAuth tokens issued to applications that are not on the approved list. Endpoint signals surface locally installed agent IDEs and personal accounts being used for work. Expense data surfaces personal AI subscriptions being reimbursed through corporate cards. Direct disclosure, framed correctly, surfaces the agents employees built but never told anyone about. None of these channels is complete on its own. All of them together start to produce an actual inventory.

The inventory has to be living, not a one-off audit. New agents get deployed weekly. Existing agents get extended with new capabilities. The discovery programme has to operate on a continuous basis, because the underlying environment is continuous. Industry guidance from the Cloud Security Alliance and similar bodies has been explicit about this: a snapshot inventory is not governance. Continuous discovery is the only model that actually works at the speed agents are appearing.[4]

For C-suite leaders, the right question is not "do we have an AI inventory?" It is whether the inventory reflects what is actually running today, how confident the organisation is in that reflection, and how recently the discovery layer was last refreshed.

Pillar 2 - Identity and Access for Non-Human Actors

Most enterprise identity infrastructure was designed around two categories of actor: human users and defined service accounts. Agents fit neither cleanly, and trying to govern them through one of those categories is where most shadow agent risk actually originates. An agent inheriting a human user's full permissions, acting on their behalf, is a different operational entity than the human, and it should be treated as one. Without that separation, there is no technical basis for distinguishing "the employee did this" from "the agent did this on the employee's behalf" in a regulatory audit, which is precisely the question regulators will ask.

The emerging consensus across major analyst houses is that agents need first-class Non-Human Identity management. Every shadow agent discovered through the inventory process should be mapped immediately to a machine-readable workload identity, distinct from the user that deployed it, with its own permissions scoped to least privilege and just-in-time access where possible. Each action the agent takes is logged against its workload identity rather than the user's, creating the technical separation that makes audit and incident response possible at the right granularity. Google Cloud's 2026 forecast describes this as agents becoming "distinct digital actors, each with its own managed identity," and recommends that organisations design identity frameworks accordingly.

This is harder than it sounds because it cuts across infrastructure most enterprises have spent a decade hardening for human users. Identity providers, SSO platforms, privileged access management tools, and access review processes were not designed for the cardinality or churn that agents introduce. The work involves either extending existing systems significantly or adopting agent-specific identity infrastructure that can interoperate with what is already in place. Either way, it is not a tooling decision that can be deferred. It is the structural foundation on which every other governance pillar rests.

Pillar 3 - Behaviour and Action Monitoring at Runtime

Even with discovery and identity in place, governance still depends on knowing what agents are actually doing once they are running. Approval at deployment is not enough. Agents drift. Their behaviour changes as the underlying models update, as the data they read changes, as new tools get connected. An agent that was safe at deployment can become risky later, and the only way to know is continuous behavioural monitoring.

This means runtime visibility into the actions agents are taking, not just whether they were approved at deployment. It means policy enforcement at the point of action, not retrospective review. It means anomaly detection that catches when an agent does something inconsistent with its declared purpose. And it means inference interdiction: the ability of the data layer to block a tool call in real time when the agent's intent does not match its sanctioned task profile. This moves the governance posture from passive monitoring after the fact to active prevention at the moment of action, which is the only model that actually works at the speed agents operate. McKinsey's research frames this as the difference between governance of intent and governance of behaviour, and it is right: most enterprise programmes have invested in the first and barely started on the second.

The hard part is that behavioural monitoring for agents requires telemetry that did not exist five years ago. It is not enough to know what the agent ran. The audit trail needs to capture what the agent saw, what context it had, what tools it considered, what alternatives it weighed, and why it took the action it took. EY and AIUC research from 2026 found that only 38% of organisations monitor AI traffic end-to-end across prompts, tool calls, and outputs. Only 17% continuously monitor agent-to-agent interactions. Those numbers will improve, but they describe the current starting point honestly: most enterprises have no real-time visibility into what their agents are doing.

Pillar 4 - Sanctioned Alternatives That Compete With Shadow Tools

This is the pillar that risk and security leaders sometimes resist because it does not look like a control. It is a control. The most reliable way to reduce shadow agent usage is to give employees a sanctioned alternative that is good enough that the unsanctioned one is not worth the friction of building and hiding.

Industry research has been consistent about this. When approved AI tools are provided that match what employees can build for themselves, unauthorised usage drops sharply. Conversely, when the sanctioned alternative is significantly worse than what employees can get for themselves with a free account, no amount of policy will keep them from using the better one. The economics are obvious. Employees are rational actors. They will use what works.[6]

This means the organisation has to be in the business of providing AI tooling that genuinely competes with the consumer offerings. Not a watered-down enterprise version that takes six months to deliver something the employee could build over a weekend. A real platform with the real capabilities, properly governed, with the data layer underneath designed to support the use cases the business actually has. This is harder than buying a license to a vendor product. It requires the data engineering investment that the previous whitepapers in this series have argued for. It is also the only sustainable answer.

"Banning backfires. Research consistently shows that nearly half of employees would continue using personal AI accounts even after an organisational ban. Prohibition drives shadow AI deeper underground rather than eliminating it."

Industry analysis, 2026

Pillar 5 - Governance Embedded in the Data Layer

The previous four pillars are necessary but not sufficient. Shadow agent governance ultimately fails or succeeds in the data layer, because that is where the agents are reading and writing. If the data layer enforces access controls, privacy rules, lineage tracking, and policy constraints automatically, an agent operating against it is governed by the layer regardless of who deployed it. If the data layer enforces nothing, an agent with the right credentials can do whatever the credentials allow.

This is where two decades of data infrastructure engineering becomes a direct competitive advantage in solving the shadow agent problem. The data layer, built correctly, acts as a universal governance broker: enforcing access controls, privacy rules, lineage tracking, and policy constraints automatically, regardless of whether the agent operating against it was formally sanctioned or deployed without anyone's knowledge. A shadow agent with the right credentials hits the same governance layer as a fully approved one. The controls do not depend on the agent having been through a procurement process. They depend on the data layer having been designed for accountability, not just availability. Lineage at the call level. Version history with point-in-time queryability. Decision logs that integrate with the data plane. Full traceability from action back to context. None of these are optional any more, and they are particularly not optional when the agents acting against the data layer were not formally deployed in the first place.

For organisations that have invested in agent-ready data layers, shadow agent governance becomes meaningfully easier, because the runtime controls already exist. For organisations that have not, the work has to happen now, because building it under regulatory pressure after an incident is significantly more expensive than building it deliberately before one.

For C-suite leaders, the right question is not whether the organisation has an AI governance committee. It is whether the data layer enforces governance automatically when an agent (sanctioned or otherwise) tries to do something it should not.

Section 5

What Effective Shadow Agent Governance Delivers

The case for investing in shadow agent governance is sometimes framed defensively: avoiding incidents, avoiding fines, avoiding the kind of reputational damage that follows when an autonomous system does something the organisation cannot explain. That framing is not wrong, but it understates the actual return. Organisations that have done this work properly are not just more defensible. They are more capable. They have unlocked the productivity benefits of agents at scale without taking on the exposure that has frustrated less prepared peers.

Visibility That Closes the Gap

The first deliverable is the most basic, and the one most organisations are still missing. A reliable answer to the question "what AI agents are operating in our environment right now?" Most CISOs cannot answer that question with confidence today. The ones who can have made the discovery investment, and the ability to answer it changes every other conversation. Risk reviews become tractable. Regulatory inquiries become manageable. Incident response becomes possible. The 68-point gap between perceived governance and actual governance starts to close, because the organisation can finally measure what it has.

Productivity Without the Exposure

Organisations with mature shadow agent governance get the productivity benefits employees have been chasing, without the exposure they have been creating. The agents move into a sanctioned environment where they can be governed, monitored, and improved. The data layer enforces the controls automatically. Employees get the tools they wanted, with proper support, without having to hide what they are doing. The organisation gets visibility, accountability, and the ability to learn across teams about which agentic patterns actually deliver value, which is impossible when half the deployment is invisible.

None of this is theoretical. Industry research consistently shows that organisations with mature AI governance frameworks see fewer security incidents, lower breach costs, and higher productivity from sanctioned tools. The pattern across the data is clear: governance done well is a productivity enabler, not a productivity tax.[16]

$670K
additional breach cost for organisations with high shadow AI exposure
(industry research, 2026)

80%
of organisations have already encountered risky agent behaviours
(McKinsey, 2025)

EUR 35M
maximum fine for prohibited AI systems under the EU AI Act
(or 7% of global turnover)

Regulatory Posture That Survives Scrutiny

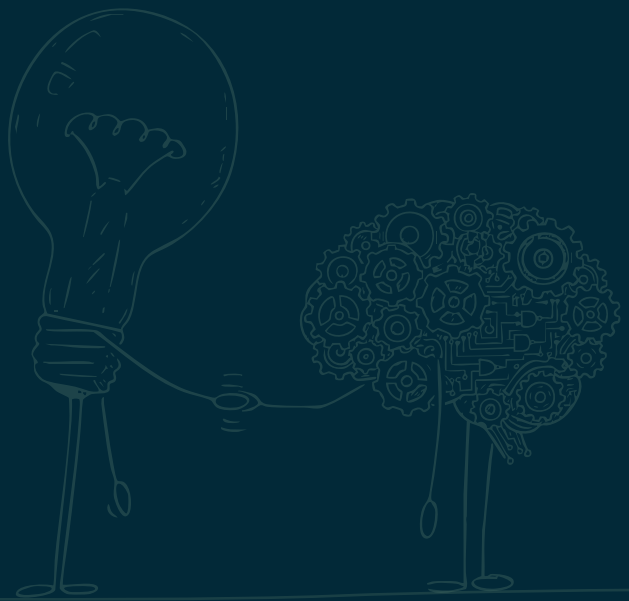
The EU AI Act, in force from August 2026 for high-risk systems, is the first of what will be a series of regulatory frameworks that assume the organisation can answer specific operational questions about its AI deployments. What agents are running. What they are doing. What they have done. What context they had when they did it. What controls were applied. Organisations that built the discovery, identity, and behavioural monitoring infrastructure before the regulation matured are positioned to answer those questions. Organisations that did not are positioned to be late, expensive, and visible to regulators in ways that no executive wants their organisation to be visible.[11]

Beyond the EU AI Act, sectoral regulators in financial services, healthcare, and energy are issuing increasingly specific guidance on AI risk management, and the pattern is similar: assumed visibility, assumed audit trail, assumed governance at the action level. The investments that make a shadow agent programme work also make these regulatory conversations easier. The organisations that have done the work will be deploying more confidently in regulated sectors while less prepared competitors slow down.

Trust That Lets the AI Strategy Actually Work

There is a softer return that does not show up cleanly in spreadsheets but matters enormously to anyone who has tried to run an enterprise AI programme through a major incident. Trust. Once a shadow agent has caused a visible problem, the organisational appetite for agentic AI tends to collapse. Boards become cautious. Risk committees become restrictive. Business units that were enthusiastic become defensive. The work of rebuilding executive confidence in AI takes years, and the leaders who were associated with the incident often do not recover their authority on the topic.

Organisations that have built proper shadow agent governance avoid this. The deployments that succeed get scaled. The ones that have problems get caught early. The board sees a programme that is being run with the same operational discipline as any other production system. The CEO can answer questions about AI strategy with credibility. The CISO is not having to defend the indefensible. The CFO can plan around a cost profile that is actually visible. The cumulative effect over two or three years is significant, and it compounds in favour of the organisations that did the work early.



Section 6

Where to Start: A Practical Roadmap

The most common mistake organisations make at the start of a shadow agent programme is treating it as a security project. It has security dimensions, but framing it that way produces a programme that is incomplete from the first design decision. Shadow agent governance is a cross-functional capability that sits across security, data, AI, legal, HR, and the business units actually deploying the agents. The CISO has to be central, but cannot drive it alone. The CDO and the head of AI have to be in the room. Legal and compliance have to be in the room. Without that breadth, the programme either stalls or solves the wrong problem.

The other common mistake is starting with policy. Policy is necessary but it is not the place to start. Policy without operational measurement is performative, and most enterprises already have policies that look fine on paper while the actual estate looks nothing like what the policy describes. The right starting point is the same as for any unfamiliar operational risk: figure out what is actually happening, then design around what you find.

Phase 1 - Discover

Before any policy gets revised or any new tool gets selected, the organisation needs an honest picture of what agents are actually deployed in its environment. This is the foundation everything else is built on, and it is the stage most organisations skip or compress because it is uncomfortable. The output of this phase is not a slide deck. It is an inventory: what agents are running, where, on whose behalf, with what permissions, against what data.

The discovery programme combines network telemetry, identity logs, endpoint signals, expense analysis, and structured employee disclosure. None of these channels is complete on its own, but together they produce an actual picture. The disclosure component is often the most productive in the early stages, because employees mostly want to keep using their agents and will tell the organisation what they are running if the conversation is framed correctly. The technical channels catch what the employees do not disclose. Both matter.

Phase 2 - Triage and Stabilise

With the inventory in hand, the next priority is triaging what was found. Not every shadow agent is a high risk. Some are reading public data and writing to nothing. Some are touching production customer data with no audit trail. The risk is not uniform, and the response should not be either.

High-risk agents need immediate attention: scope reduction, identity separation, monitoring deployment, and in some cases shutdown until proper controls are in place. Medium-risk agents need a path to sanction through a sanctification pipeline: the agent moves into a secure sandbox where the code is reviewed, credentials are rotated away from the employee's personal tokens, and a workload identity replaces the inherited permissions. The employee keeps using it, the organisation gets the audit trail and the controls, and critically the process rewards transparency rather than punishing it. Employees who disclosed their agents voluntarily see them preserved and improved rather than shut down, which sets the cultural tone for everything that follows. Low-risk agents may simply need to be acknowledged in the inventory and revisited later.

Stabilisation also includes the foundational identity work. If agents have been running with employee credentials, the path forward involves giving them their own managed identities with least-privilege scopes. This is non-trivial engineering, particularly at scale, but it is where the actual governance starts to bite.

Step 3 - Build Sanctioned Alternatives

Discovery and triage solve the visibility problem. They do not solve the underlying demand. Employees built agents because they needed agents, and that need does not disappear when the shadow ones get governed. The third phase is about giving employees sanctioned tools that genuinely compete with what they could build for themselves. Not a watered-down version. The real thing, with the real capabilities, properly governed, available without a six-month procurement process.

This is where shadow agent governance connects back to the agent-ready data layer the previous whitepaper in this series argued for. The sanctioned alternative needs the streaming, memory, semantic, and governance infrastructure to actually work for the use cases employees care about. Without that, the sanctioned alternative is worse than the shadow option, and employees will rationally choose the shadow option. The data layer investment is what makes the governance programme sustainable.

Step 4 - Embed Continuous Governance

The final phase is about making the controls live continuously rather than being a one-off cleanup. Discovery becomes a continuous capability, not a quarterly audit. Identity provisioning for new agents becomes automated rather than ticket-driven. Behavioural monitoring runs in production with real-time alerting. Sanctioned alternatives get extended as new use cases emerge, so employees do not have to choose between the sanctioned tool and the productive one. The governance programme becomes part of how the organisation operates rather than a parallel workstream.

This is where the compounding starts. Each new agent deployed on the sanctioned platform inherits the controls. Each new use case adds value to the data layer rather than creating a new shadow problem. The governance gap closes as the runtime infrastructure matures, not as the policy document gets longer. Over two to three years, organisations that have done this work properly are operating with substantially more visibility, less exposure, and more productivity from AI than peers still trying to ban their way out of the problem.

How Merit Approaches This

Merit's approach to shadow agent governance starts where it should start: with discovery and an honest inventory. We do not begin with policy templates or framework slides, because they tend to produce the appearance of governance rather than the substance of it. We work alongside CISO, CDO, and AI leadership teams to build the runtime visibility that lets the organisation see what is actually deployed, then sequence the identity, monitoring, sanctioned alternative, and data layer investments that turn visibility into governance. The output is not a policy document. It is an operating model that scales with the agent estate as it grows.

Our clients do not inherit a brittle compliance framework from us. They inherit a governance programme that holds up under regulatory scrutiny, supports the productivity their employees want, and gives the C-suite a defensible answer to the questions boards are increasingly asking about how the organisation governs its autonomous systems.

Conclusion

The question facing every CISO, CRO, CIO, and CEO with an AI agenda in 2026 is not whether shadow agents are running in their environment. They are. The question is whether the organisation can see them, govern them, and respond to them before something visible goes wrong.

The 40% cancellation forecast that has put a quiet shadow over agentic AI investment is not just about formally deployed agents. It is also, increasingly, about the agents nobody approved that were running anyway when a regulator, a customer, or an audit team came asking questions. The organisations that get through 2027 in good shape are the ones that started building visibility, identity, monitoring, and sanctioned alternatives in 2026. The ones that did not will be explaining to their boards why a system they did not officially deploy did something they cannot officially explain.

This is not a problem that policy alone will solve. It is not a problem that bans will solve. It is a problem that requires the same operational discipline applied to any production system: visibility, accountability, control, and a deliberate plan for when things go wrong. The work compounds. The organisations that started early have a structural advantage over the ones starting late. The ones that have not started yet are running on borrowed time, and the time runs out faster every quarter.

2026 is not the moment to write the shadow agent policy. It is the moment to find out what is actually running in the environment, and start building the governance that makes the answer manageable.

Ready to find out what is actually running in your environment?

Merit's data and AI teams work with CISO, CRO, CIO, and CDO leadership to design and operate shadow agent governance programmes that scale with the agent estate. We work from discovery and triage through to sanctioned alternatives and continuous runtime governance, ensuring the organisation can see, govern, and explain what its autonomous systems are doing.

meritdata-tech.com/ai

About Merit Data & Technology

Merit Data and Technology is part of Merit Group PLC. For over 20 years, we have been helping enterprises design and build data infrastructure that performs under real-world conditions, not just in the proof of concept but in production, at scale, and under regulatory scrutiny.

Two decades in this field means we have seen technology cycles come and go. We have worked with organisations that got the architectural decisions right early and compounded the advantage, and with organisations that deferred those decisions and paid a steep price to recover. That experience shapes how we approach every engagement: with the rigour, the realism, and the long-term perspective that high-stakes data and AI decisions deserve.

For C-suite leaders who need more than a vendor, who need a partner with the track record to be trusted with decisions that will define their organisation's AI trajectory for years to come, Merit brings 20 years of delivery credibility to the table.

Sources

1. Google Cloud (2025). Google Cloud Cybersecurity Forecast 2026. Analysis of agentic identity management and shadow agent risks for enterprises. cloud.google.com/security/resources/cybersecurity-forecast

2. McKinsey & Company (2025). Deploying Agentic AI with Safety and Security: A Playbook for Technology Leaders. McKinsey Quarterly, October 2025. Finding: 80% of organisations have already encountered risky agent behaviours. mckinsey.com/quarterly/overview

3. Gartner (2025). Press Release: Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027. June 25, 2025. gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

4. Cloud Security Alliance (2026). Research Note: AI Agent Governance Framework Gap. AI Controls Matrix (AICM) covering 18 security domains and over 240 control objectives mapped to the AI lifecycle. cloudsecurityalliance.org/research

5. Gartner (2025). Forecast cited in press release: 40% of enterprise applications will embed task-specific AI agents by end of 2026, up from less than 5% in 2025. Ibid.

6. EY (2026). Technology Pulse Poll. Survey of 500 US technology executives. Finding: 52% of department-level AI initiatives operating without formal approval; 45% of technology executives confirmed or suspected sensitive data leaks from employees using unauthorised generative AI tools. ey.com/en_us/insights/technology

7. ISACA (2025). The Rise of Shadow AI: Auditing Unauthorized AI Tools in the Enterprise. Industry guidance on AI usage audits and discovery frameworks. isaca.org/resources/news-and-trends/industry-news/2025/the-rise-of-shadow-ai-auditing-unauthorized-ai-tools-in-the-enterprise

8. Gravitee (2026). State of AI Agents Report. Finding: only 14.4% of organisations have full security approval for all AI agents currently deployed. gravitee.io/blog

9. McKinsey & Company (2026). Trust in the Age of Agents: Agentic AI Governance for Autonomous Systems. mckinsey.com/capabilities/risk-and-resilience/our-insights/trust-in-the-age-of-agents

10. McKinsey & Company (2026). State of AI Trust in 2026: Shifting to the Agentic Era. AI Trust Maturity Survey of approximately 500 organisations. mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era

11. European Union (2024). Regulation (EU) 2024/1689 – Artificial Intelligence Act. High-risk AI obligations entering full force August 2026. eur-lex.europa.eu/eli/reg/2024/1689/oj

12. Deloitte (2025). Unpacking the EU AI Act: The Future of AI Governance. Analysis of penalty structures and compliance requirements for high-risk AI systems. deloitte.com/us/en/services/consulting/articles/eu-ai-act-ai-governance.html

13. BlackFog (2026). Shadow AI Enterprise Survey. Research finding that approximately half of employees would continue using personal AI accounts after an organisational ban. Cited in CIO, January 2026. cio.com/article/4124760/roughly-half-of-employees-are-using-unsanctioned-ai-tools-and-enterprise-leaders-are-major-culprits.html

14. Deloitte AI Institute (2026). State of AI in the Enterprise 2026. Survey of 3,235 senior leaders across 24 countries. Finding: only one in five companies has a mature governance model for autonomous AI agents. deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html

15. EY / AIUC-1 Consortium (2026). Survey published in Help Net Security, March 2026. Finding: only 38% of organisations monitor AI traffic end-to-end across prompts, tool calls, and outputs; 17% continuously monitor agent-to-agent interactions. helpnetsecurity.com

16. Deloitte (2025). AI in the Boardroom: Five Governance Actions. Analysis of board-level AI governance practices. deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/articles/ai-boardroom-governance-five-actions.html