

A woman with dark hair and glasses, wearing a black top and white trousers, stands in front of a large wooden door. She is holding a laptop and a smartphone, and wearing a lanyard with a badge. The background is slightly blurred, showing other people and a wooden structure.

Guide



"On the internet, nobody knows you're a dog"

A practical guide to
starting small, reducing
risk and building secure
foundations before scaling
agentic identities

hippodigital.co.uk

Agentic AI can help organisations improve customer experience, reduce manual work and respond faster to operational pressure. But when AI systems can use tools, access business data and take action across workflows, they also create new cyber, identity and supply chain risks.

This is a practical exploration of implementing low risk use cases, treating agents as identities and applying established, foundational cyber security controls from the outset.

What this guide covers

Agentic AI implementations have become operational for business delivery.

AI agents enable autonomous action, shortening the decision making chain which leads to increased outputs. However, as agents are often given unmanaged and widespread access to organisational data and tools, their use cannot be treated as a normal automation project. Once an agent can use tools, access data and take action across systems, it must sit within an organisation's cyber security model.

- Why agentic AI needs a different security conversation to that around 'standard' generative AI
- How to choose low risk use cases before autonomy scales
- Associated risks of non-human identities, static credentials and secret sprawl
- How agent speed and unmonitored behaviours create new cyber risk
- What controls should be in place before agentic AI is connected to higher impact systems
- Questions leaders should ask before scaling a pilot

Introduction

Agentic AI is moving rapidly from experimentation into practical business use.

The use cases are easy to see. Agents can help teams summarise information, support customer service, triage operational issues, monitor risk, connect data across systems and reduce repetitive work. In all areas, carefully chosen use cases could improve both internal delivery and external customer experience.

A generative AI tool might draft a response, summarise a policy or suggest a next step. An agentic AI system is designed to go further. It can interpret a goal, plan the steps required, call tools, access APIs, use business data and carry context across a workflow. In some cases, it may also update records, trigger processes or interact with supplier systems.

Agent use therefore carries its own security risk. An agent's ability to act and interact creates an environment which requires organisations to take the next level of care.

If an agent has access to customer data, operational systems, supplier platforms or internal tools, it becomes a privileged user in its own right. It needs permissions, monitoring, boundaries and a named authoriser. It also needs to sit within the organisation's wider approach to identity, supply chain management and incident response.

The safest starting point is a controlled, low risk use case. One where the task is narrow, the data is approved, the impact of error is limited and a human can review the output before anything significant happens.

For example, a cyber operations agent can be used to review low security alerts and retrieve additional context, as well as correlate recent or related threat activity associated with that asset or identity. With read-only access, the agents can't initiate automations or actions and human analysts remain responsible for any final decisions. This shows how an agent could use reasoning across several tasks, whilst any automation elements can be constrained.

Early pilots like this should test the security model as well as the business case. Before an agent is connected to higher impact systems, teams need to know whether they can monitor it, limit its access, understand what it has done and remove access quickly if something changes.

A low risk pilot should prove the controls as much as the use case.



Glossary of key terms

Term	What it means
Agentic AI	Goal driven AI systems that can maintain context, plan steps, use tools and carry out actions across a workflow.
Non-human identity, or NHI	A machine identity such as a service account, API key, token or certificate that allows a system, workload or agent to access another system.
Zero trust	A security approach that assumes no implicit trust. Identities, devices, workloads and requests are continuously verified and access is limited to what is needed.
Least privilege	Giving a user, system or agent only the minimum access needed to complete an approved task.
Human in the loop (HITL)	A governance model where people review, approve or intervene in AI supported decisions, especially where the risk or impact is high.
Supply chain risk	The risk created by suppliers, platforms, tools, integrations, models and third party services connected to an organisation's systems or data.
NCSC	The UK's National Cyber Security Centre, which provides guidance to organisations on cyber security and governance.
DORA	The EU's Digital Operational Resilience Act. It applies to financial entities, but is useful as an example of how cyber resilience, incident reporting and third party risk expectations are becoming more formal.
Workload identity	A way of assigning a verifiable identity to a workload, service or agent so it can access systems without relying on shared or long-lived credentials.
Static credentials	Long-lived, unchanging authentication tokens such as API keys, service accounts or certificates that act as a permanent open door into connected systems.
Secret sprawl	The uncontrolled and unmonitored spread of machine credentials across code repositories, CI/CD logs, messaging platforms, collaboration tools, and cloud-stored spreadsheets.
Time-to-live (TTL)	Refers to the expiration period assigned to an authentication credential, token, or session. Once the TTL expires, the credential automatically becomes invalid, requiring renewal or re-authentication.

Part one

Is an agent needed at all?

Before deploying an agent, organisations should check whether agentic AI is the right tool for the task.

Some problems do not need autonomy. They may need clearer ownership, cleaner data, better process design or a standard automation. If the underlying workflow is already unclear, adding an agent can make it harder to manage.

Agentic AI is most useful where a task genuinely needs reasoning across several steps, use of tools or adaptation to context. It is less useful and increases risk where a search function, rules based workflow or normal integration would solve the problem.

Prior to any pilot or agent implementation, a roadmap should include a discovery step. That would include assessing:

- > What problem are we trying to solve?
- > Does the task genuinely require an agent?
- > Could the process be simplified first?
- > What AI services, non-human identities or agents are already in operation?
- > Which data and systems need to be accessed by and exposed to agent activity?
- > Whether workload identity and least privilege can be enforced
- > What actions the agent needs to be able to take, and whether these can be logged, tracked and correlated
- > Whether there is a mechanism to stop the agent or revoke access
- > Who owns the agent and is responsible for its risk?
- > What the threat model looks like
- > What is the worst realistic outcome if it behaves unexpectedly?

If these questions cannot be answered clearly, the use case is not ready.



Why agentic AI should start with low risk tasks

Use cases must be practical.

Agent deployment should offer clear value that outweighs both its financial cost and operational risk. Unlike standard generative AI systems which may leak sensitive data in chat interfaces but cannot alter core business systems, a credentialed agent operating across enterprise environments has the potential to pose a far greater identity and cyber security threat. Consequently, Agentic AI governance demands more rigorous monitoring techniques through existent cyber security controls.

Low risk use cases give teams a safer place to test the technology and the controls around it. A pilot should show whether the agent creates value, but it should also show whether the organisation can see what the agent did, limit where it can go, review its output and remove access if something goes wrong.

Good early use cases tend to be internal, narrow and easy to supervise. They should use approved data sources, have a clear owner and have limited impact if the output is not the expected one.

Examples could include:

- > Summarising internal documents
- > Searching an approved knowledge base
- > Drafting internal status updates
- > Preparing meeting notes from approved material
- > Triage of low severity service desk tickets
- > Flagging missing information in forms
- > Helping colleagues find the right internal process
- > Producing first draft responses for human review

These use cases can still reduce manual effort and improve access to information, without giving an agent broad access to customer records, production systems, payment flows or supplier platforms.

Higher risk use cases should come later. These might include changing customer records, approving refunds, updating supplier data, modifying access permissions, interacting with production systems or handling sensitive personal data.

A useful pilot gives teams evidence. It shows whether the agent can stay within scope, if the logs are good enough, if people understand when to intervene and whether the access model is strong enough for the next stage.

Low risk versus higher risk starting points

Lower risk starting point	Made safe by	Higher risk use case	Why it needs more human governance
Summarising internal policy documents	Uses approved content and can easily be reviewed before use. Scope can be clearly defined and limited	Updating customer records	Could affect customer service, record accuracy and audit trails
Searching a knowledge base	Limited to existing guidance. Agent access should be read only and sources restricted	Approving refunds or compensation	May affect money, customer trust and fraud controls
Drafting internal updates	Human must review before sharing. An agent should be made unable to automatically publish updates and data classification/DLP controls are applied to ensure confidential information is not unintentionally incorporated into a draft	Changing access permissions	Could expose systems or data if misconfigured
Triage of low severity tickets	Helps route work, classify and enrich tickets without final decision-making authority	Acting on production systems	Could affect service availability and operational resilience
Flagging missing information in forms	Supports human review	Interacting with supplier systems	Creates third party and supply chain exposure



Applying established controls from the outset

To implement these use cases, several controls should be established from the outset, and which are implemented relates to the level of risk an organisation is willing to accept. Agentic AI may be a more recent and less mature area of delivery, but the starting point should be based on established cyber security practices as well as People, Process and Technology principles.

People, Process and Technology principles for AI agents

Principle	Agentic AI Consideration
People	Named owner/sponsor; competent human approvers; segregation of duties; AI/security awareness; incident authority; accountability for deployment and permissions
Process	Agent registration; risk assessment; threat modelling; approval; access reviews; change management; security testing; incident response; decommissioning; recurrent evaluation cadences, performance reviews and remediation workflows for model drift
Technology	Least privilege; agent identity; short lived credentials; policy enforcement; API controls; data controls; secrets management; automated monitoring tools, telemetry and metrics pipelines for evaluation and model drift detection.

Agentic AI risks and mitigations should be aligned with the organisation’s existing security model and risk posture. The controls should be designed into the pilot before it begins, rather than added after the agent has already been connected to data or systems.

Baseline controls for early agentic AI pilots

Control	What it means in practice	Why it matters
Assign unique identifier and owner	Every agent has a dedicated identity plus a named human. No shared human credentials	Ensures accountability and traceability of each agent
Apply least privilege	Give agents only the access they need, for the shortest time required	Limits the impact of error or compromise
Limit scope	Define what the agent can access, what it can do and when it must stop	Reduces ambiguity and unexpected behaviour
Avoid long lived credentials	Use temporary credentials where possible and revoke elevated access after the task	Reduces the exposure window if credentials leak
Manage full identity lifecycle	Provision, review, change and retire the identity alongside the agent. Access should not survive the workload	Ensures NHIs do not become dormant or orphaned and are decommissioned as and when appropriate
Use secure defaults	Start from restricted access, safe configuration and validated inputs	Prevents temporary pilot shortcuts becoming permanent weaknesses
Understand dependencies	Map models, tools, integrations, suppliers and data flows	Makes supply chain exposure visible before deployment
Ensure audit trail and monitor behaviours	Capture audit data from identity, actions, tools, target resources, approvals and outcomes. Enables early detection of unusual or unexpected activity across tools and connected systems	Helps teams spot risk early
Treat inputs and tool outputs as untrusted	Protect against direct / indirect prompt injection, tool poisoning and malicious retrieved content	Ensures the agent remains bounded, secure and resilient against malicious manipulation
Threat model the deployment	Consider misuse, manipulation, excessive access, prompt injection and supplier compromise	Tests the design before live use
Plan for incidents	Define how to suspend the agent, revoke access, preserve logs and investigate	Reduces confusion and uncertainty when something goes wrong



Human in the loop governance

Another governance tool is the human in the loop (HITL) model. The concept is predicated on the idea that the humans and the approval processes all need to be secured. Humans should still be authenticated and authorised to approve actions, ideally in a trusted approval channel, and all approvals should be recorded.

Agents should never be able to choose an approver.

Humans should still be subject to existing least privilege and zero trust principles.

A human in the loop model therefore allows the agent to gather information, summarise evidence and recommend a next step, but requires that a person is the final authorisation for an action.

This is useful for:

- > Refunds and compensation
- > Complaint handling
- > Account restrictions
- > Fraud investigation
- > Service eligibility decisions
- > Supplier changes
- > Access permission changes
- > Production system updates
- > Incident response actions

An agent's governance model should account for the risk of the task. A lower risk activity (such as producing communications based on internal policy) may only need active behaviour monitoring, review and final approval by a human in the loop.

Conversely, decisions that may affect customers, their money, safety, access or service continuity are high risk activities that require closer monitoring and possibly even individual approval at every step.

Necessary governance is defined by the use case. A human in the loop governance model protects your business and your customers whilst allowing for innovation.

Without a clear monitoring strategy and internal governance in place targeting agent creation and scope, security teams are looking at a dark space where activity is occurring, but without human knowledge or intervention.

Possible risks include;

- > Increasing difficulty of managing non-human identities
- > Secret sprawl and static credentials
- > Use of shorter lived credentials
- > Vulnerabilities caused by agent implementations
- > Insufficient control of unexpected behaviour

Identity as the control point

Identity is where technical control and accountability meet.

Agents originate internally, often gaining privileged access at inception. They may operate across cloud services, internal systems, supplier platforms, models and data stores. They may not sit neatly inside one network boundary. As workflows become more connected, an identity based security model becomes a more reliable control.

Every agent should have its own identity. It should never share a person's credentials. It should not inherit broad access by default. It should not keep permissions indefinitely.

The organisation should be able to answer:

- > Which agent acted?
- > Under whose authority?
- > What data did it access?
- > Which tool did it call?
- > What was it allowed to do?
- > Was the action approved?
- > Can access be revoked quickly?

This is especially important for customer facing and operational use cases. If an agent supports a refund, complaint, account change, service update or supplier interaction, the organisation needs to understand and explain what happened. There must be clearly defined agent monitoring in place, a human in the loop: a person accountable for the agent's behaviour. This creates accountability that underwrites agent activity.

An agent Joiner-Mover-Leaver (JML) process manages what agents can access when; what the current scope for that identity is, and it manages the decommission stage. It shows what acted, what it accessed and what authority it used. This necessitates a really tight mover process in particular, as agents may change roles, scope and permission needs more often and with more lateral movement than a human user.

Agents should not share human credentials. Each agent needs its own identity, access boundary and audit trail.



As we have seen, the AI agent is a new type of digital identity. It is not just a user, but a user with a specific role and a specific set of permissions. This means that the AI agent is a new type of digital identity that needs to be managed and controlled. The AI agent is a new type of digital identity that needs to be managed and controlled. The AI agent is a new type of digital identity that needs to be managed and controlled.

An additional dimension to identity comes in the form of supply chain scrutiny. Suppliers of AI agents need due diligence applied to their onboarding as there are often transient dependencies in agents that must be surfaced and understood. In addition to the questions above before deploying an agent, organisations should understand:

- > Which suppliers are involved
- > Which systems the agent will connect to
- > What data will pass through those systems
- > Where that data will be stored
- > Which supplier controls the platform
- > What logs the supplier can provide
- > How quickly access can be revoked
- > Who is responsible if the agent contributes to an incident
- > Whether the supplier’s controls meet the organisation’s standards

This is particularly important for retail and customer facing businesses. If an agent connects to order management, loyalty, payments, logistics or customer service platforms, supplier risk becomes part of the AI risk model.

Building secure autonomy: a practical roadmap

With this wider context in mind, organisations do not need to wait for a perfect strategy before employing agentic AI. They do need a governed route into adoption.

Adoption roadmap

Stage	What to do	Evidence to capture
1. Begin a discovery and assessment	Map existing AI use, non-human identities, secrets, service accounts, API keys and supplier integrations	Known machine identities, exposed secrets, over privileged accounts and dormant access
2. Select a low risk pilot	Choose a narrow, internal use case with a named owner and approved data sources	Use case brief, risk assessment and success criteria
3. Modernise the identity layer	Move away from static secrets where possible and use short lived credentials, workload identity and real time policy checks	Access model, credential lifecycle and revocation process
4. Build monitoring into the workflow	Track agent activity across tools, data sources and connected systems	Normal behaviour profile, alert triggers and review process
5. Design for accountability	Define ownership, approval, monitoring, incident review and stop authority	Responsibility model and human review points
6. Align with zero trust principles	Verify identities and behaviour continuously and apply least privilege to data, tools, models and actions	Policy checks, access logs and audit trails
7. Review before scaling	Decide whether to expand, adapt or stop the pilot based on evidence	Performance, control effectiveness and unresolved risks
8. Explore rollback	Plan your contingency, ensuring you can cleanly revoke privileges and reverse actions	Implement an active agent monitoring process and ensure ongoing production of decision logs



Part two

Navigating risk: problem mapping

All agentic activities carry some element of risk, and businesses mitigate that cyber risk in several ways. Human error, unnecessary data access through generative models and over-credentialed agents can all lead to data breaches.

Here is an outline of some of the risks associated with agent implementation and investigations you can do with current monitoring techniques to ensure you can manage agent activity across systems.

This comes down to the need for a stronger control model for agentic AI, as the risk is not only in the model output. It is in the access, actions and connected systems around the agent.

Visualising the difference: access models compared

Model	How access works	Main risk
Human user	A person logs in, takes action and is reviewed through normal access controls	Phishing, social engineering, user error, excessive access or weak offboarding
Generative AI copilot	A user prompts a system and decides what to do with the output	Data leakage, inaccurate output, poor review or unsafe use of sensitive information
Agentic AI agent	A system interprets a goal, uses tools, accesses data and acts across a workflow	Credential sprawl, excessive permissions, unexpected behaviour, supplier exposure and actions happening faster than humans can review

Problem one

Why non-human identities are becoming harder to manage

Agentic systems rely on tool use. Each API call, database query or microservice invocation requires some form of proof of identity or machine credential for the agent.

That credential might be a service account, API key, token or certificate. These are not new, but agentic AI significantly increases their critical importance, as an agent often needs to interface with multiple systems to execute a single task where each connection demands its own credentials and access requests.

Research from Entro Labs’ H1 2025 NHI and Secrets Risk Report found that 7.5% of machine identities remain active for five to ten years. It also found that 1 in 20 AWS machine identities carry full administrative privileges, creating what the report describes as “super-NHIs” with the ability to control entire cloud environments.

This generates a significant business risk, as many identity processes are designed around people. A person joins the organisation, gets access, may change roles and eventually leaves. There are usually approval routes, access reviews and offboarding steps in place to manage the data, tools and resources they have access to.

Conversely, non-human identities are not created and enabled via HR systems. Though many will have their own unique identifiers, they tend to be built in decentralised, ad-hoc, or fully automated ways; no ticket raised, no access approval route triggered, no offboarding date set.



Evidence point	What it helps prove	Proposed investigative steps and solution
Non-human identities outnumber human identities by 144 to 1 in large enterprises	Machine identity is already difficult to govern before scaling of agentic AI	Inventory NHI's, assign owners and maintain a central register
The ratio increased from 92 to 1 to 144 to 1 in one year	The problem is growing quickly	Track NHI growth trends and enforce policy guardrails
7.5% of machine identities remain active for five to ten years	Some identities outlive the workloads they were created for. And it's likely that, in that time, the human responsibility chain has been disrupted.	Review agent age, usage and workload dependencies, disable stale identities, and introduce automated expiry
1 in 20 AWS machine identities carry full administrative privileges	A small number of identities can create significant exposure	Identify and review privileged NHI's and reduce access using least privilege and JIT elevation
43% of leaked secrets appear in CI/CD logs, messaging platforms and collaboration tools	Secrets are not only a code repository problem	Scan pipeline tools, revoke exposed secrets and move to managed credentials
5.5% of AWS identities with admin permissions show no activity in 90 days	Dormant privileged identities can be missed by normal review processes and oversight is lost	Disable dormant privileged identities and require reapproval
A typical AWS role touches 43 different services	One compromised identity can create lateral movement risk	Map all access points and limit access

Shadow agents are created in much the same way, also bypassing HR systems and cyber and identity processes. They may be shared across systems, remaining active long after they are needed, and they may hold more access rights than the task requires.

Whilst both agents pose a risk, the risk from shadow agents is significantly higher, as they are authorised (intentionally or not) to borrow a human identity to gain authentications and authorisations, making them virtually invisible to cyber security monitoring.

This creates a significant practical problem for security teams.

A large number of potentially unmonitored, newly created synthetic users with potentially unlimited access are generating activity in the organisation's systems, and some will have the ability to change and access fundamental infrastructure.

Identity management and monitoring from inception

Security teams therefore need to manage non-human identities throughout their lifecycle. They can control agent sprawl by discovering and registering non-human identities, assigning each agent a unique identity and accountable owner, governing its permissions throughout its lifecycle, and monitoring its behaviour across the systems it accesses.

Identity, authentication and activity logs should then be monitored to detect unusual behaviour, with regular access reviews and the ability to revoke credentials or disable the identity quickly when it is no longer required.

Because, where identities are not visible, attributable, reviewed and revoked when no longer needed, agents can create persistent and difficult to detect routes into business systems.



Problem two

Secret sprawl and static credentials

Without clear governance and strategy, the credentials that authenticate non-human identities can spread across an organisation quickly.

They may appear in code repositories, CI/CD logs, messaging platforms, collaboration tools or spreadsheets stored in cloud drives. Slack bot tokens, Jira keys, API keys and certificates can become valuable targets when they are easy to find and hard to monitor.

In the same report, Entro Labs found that 43% of leaked secrets now appear in CI/CD logs, messaging platforms and collaboration tools. That matters because collaboration spaces are thought to be secure by users and often sit outside the places security teams check first.

Static secrets create a particular risk. An API key or certificate with a long lifetime can act as a persistent route into connected systems if it is exposed. The longer such a credential exists, the more time there is for it to be copied, forgotten, reused or compromised. This is particularly risky where people are employing shadow agents which use their personal identity to access systems within an organisation.

The Entro Labs research also found that nearly half of enterprise non-human identities are over a year old. It also found that 5.5% of AWS identities with admin permissions had shown no activity in 90 days. Dormant credentials with high levels of access are exactly the kind of risk that can sit unnoticed until a wider incident brings it to light.

Agentic AI makes this more difficult to manage because agents often need access to several connected tools. If those connections rely on long lived credentials, or an agent is over-privileged from inception, the agent can increase the potential impact of a breach.

Agentic workloads can fall outside controls designed primarily for interactive human users. They may operate in cloud or SaaS environments where host based EDR is not the primary source of telemetry, authenticate non-interactively through workload identities or tokens, and act across multiple services at machine speed. Monitoring therefore needs to combine identity, cloud/API, application and agent telemetry. Without clear ownership, lifecycle controls and correlated logging, over privileged or long lived agent identities can become difficult to attribute, contain and retire.

Tightly scoped access provision

The safer pattern is temporary, tightly scoped access. Agents should receive only the access needed for a specific task, for the shortest practical time. This reduces risk associated with reduced ability to monitor. Elevated access should be removed when the task is complete.

Problem three

Vulnerabilities from the inside

Agentic AI can also create risk without an external attacker.

An agent may interpret a goal in a way the user did not expect. It may take a technically valid route that creates business risk. It may use a tool outside the intended process. It may retrieve more data than needed. It may act quickly enough that a small issue becomes harder to contain.

Monitoring therefore needs to cover behaviour, not only known threats.

Define your normal

Organisations should define what normal activity looks like for each agent. They should monitor which systems it accesses, what tools it calls, how often it acts and whether it stays within the agreed workflow.

Signals worth tracking include:

- Attempts to access unapproved data
- Tool calls outside the agreed process
- Unusual frequency of actions
- Repeated failed access requests
- Requests for elevated permissions
- Outputs that conflict with policy
- Unexpected interaction with supplier systems
- Activity outside agreed time periods

This does not make agent behaviour perfectly predictable. It gives teams a way to spot activity that falls outside the intended pattern and investigate before the risk spreads.



Problem four

Using shorter lived credentials and workload identity

Shorter lived credentials reduce risk because they limit the time an attacker can use them if they are exposed. They are a valuable tool for use with discrete actions, in high risk workflows, or for automated workload scaling. Using the SPIFFE framework, credentials can be automatically issued, rotated and revoked without manual intervention.

While effective, this approach remains relatively crude.

For many projects, including deep research tasks, a longer term approach would be to implement fine-grained authorisation techniques, such as Attribute-Based Access Control (ABAC). This enables conditional access rules that allow for longer-lived credentials or adaptable refresh policies tailored to specific use cases, while strictly limiting access to relevant tasks and processes.

Minimum viable permissions and managing scope

Agents should therefore be granted minimum viable permissions for current operations. You won't be able to guess the credentials needed over the next several months, or even years, so expect to introduce additional privileges through governed change or JIT approval as the use case develops.

Consider the following when scoping:

- What stable identity represents the workload?
- What secrets / credentials are currently stored?
- What token TTLs are being used?
- How are credentials renewed?
- Can access be revoked immediately?
- What does the agent really use versus what it has been granted?
- What happens if the token issuer / IdP is unavailable?
- How does the permission set change as the agent's role changes?

Problem five

Why speed and unexpected behaviour need to be controlled early

AI can help attackers move faster. It can support reconnaissance, vulnerability discovery, phishing, credential harvesting and data extraction. Where agentic tools are used, more of that activity can be planned and executed at speed.

In November 2025, security researchers documented an AI orchestrated cyber espionage campaign in which a scaffolded version of Claude Code acted as the execution engine. According to the original case study, the agent performed reconnaissance, vulnerability discovery, exploitation, credential harvesting and data exfiltration, conducting 80% to 90% of the labour.

That figure should not be read as a sign that human attackers have disappeared. Rather, they are now AI-enabled. This shows how much of the work can now be accelerated.

Speed matters because it reduces the time available for human teams to spot and respond to problems. If an attacker can probe systems, test credentials or move between services faster than a team can review alerts, human in the loop agent management approaches fail, as speed is key. Organisations must have robust detection, monitoring and containment capabilities in place, ready for such an attack.

Defensive AI and pattern of life monitoring

Agentic AI also has a defensive role.

Security teams already manage high volumes of alerts, logs and signals. As attacker activity becomes faster and more automated, human review alone can struggle to keep pace. Defensive AI can help teams identify unusual activity, group related alerts and prioritise the events that need attention first.

Pattern of life monitoring is one useful example. A system learns what normal activity looks like for users, devices, applications or workloads. When behaviour changes, such as a sudden increase in data transfer or an unexpected series of API calls, the system can flag, throttle or isolate the activity for review.

For agentic AI, this kind of monitoring is useful because it can help organisations track unexpected behaviour. An agent that begins calling new tools, accessing additional, unconnected data or acting outside its normal pattern should be treated as a signal for investigation.

These tools should support security teams with clearer signals and faster routes to containment, while people remain responsible for judgement, escalation and response.



Conclusion

What leaders should ask before scaling

If these questions cannot be answered clearly, the agent should not be scaled.

Question	Why it matters
What is the agent designed to do?	Confirms the task is defined rather than open ended
Why is an agent needed for this task?	Tests whether a simpler approach would work
What data does the agent access?	Identifies privacy, commercial and operational risk
What systems can it connect to?	Shows where the agent sits in the technology environment
What actions can it take?	Clarifies the level of autonomy
What can it not do?	Makes the boundary visible
Who owns the agent?	Creates accountability
Who approved its access?	Shows whether governance is in place
How is the agent monitored?	Tests whether behaviour can be tracked
What behaviour would trigger an alert?	Defines what unusual activity looks like
What actions require human approval?	Keeps judgement in high impact decisions
How quickly can access be revoked?	Shows whether the organisation can contain a problem
What suppliers are involved?	Identifies supply chain exposure
What logs are available?	Supports auditability and incident response
What is the incident response plan?	Confirms the organisation is ready for failure or misuse
What evidence do we have that the controls work?	Prevents scaling based only on productivity claims



Building a secure foundation for agentic AI

Agentic AI can help organisations improve customer experience, reduce manual work and strengthen operational delivery. Used carefully, it can support better service, faster triage and more effective use of business data.

That potential depends on control. Once an agent can access systems, use tools and take action, it becomes part of the organisation's cyber security model. It needs a clear identity, limited permissions, monitoring, human oversight and an incident response plan.

The first deployments should be narrow enough to learn from and controlled enough to stop. They should prove whether the organisation can govern the agent before it is connected to higher impact systems.

For leaders, the decision is not simply whether agentic AI can create value. It is whether the organisation can explain what the agent is doing, limit what it can access and respond quickly if something goes wrong.

That is the basis for safer adoption: low risk tasks first, establish cyber security controls early, ensure clear ownership, visible dependencies and people accountable for meaningful decisions.

Source list

- NCSC: Thinking carefully before adopting agentic AI
- NCSC and international partners: Careful adoption of agentic AI services
- Entro Labs: NHI and Secrets Risk Report, H1 2025
- Microsoft Security: What are non-human identities?
- SPIFFE: Secure Production Identity Framework for Everyone
- Anthropic: Disrupting the first reported AI orchestrated cyber espionage campaign
- GOV.UK: Cyber Security and Resilience Bill policy statement
- EIOPA: Digital Operational Resilience Act overview
- Reuters: M&S pauses online orders after cyber incident



Designed for humans. Built for intelligence.

We design and deliver AI-enabled digital services where human insight, data and engineering work as one, creating systems that learn, adapt and deliver measurable impact

A trusted partner for the UK's most critical digital infrastructure



Department
for Education



Department
for Work &
Pensions



Ministry
of Defence



Department
of Health &
Social Care



Ministry
of Justice



Home Office

