Token Security

THE AI SECURITY GUIDE

A MATURITY MODEL FOR SECURE AGENTIC AI ADOPTION

token

Table of Contents

Introduction

Rapid adoption of AI technology is a crucial business initiative. In fact, 82 percent of executives at large enterprises plan to integrate AI agents within the next three years. As organizations race to adopt agentic AI, many overlook a critical factor: security. Rushing into deployment without proper safeguards can introduce serious risks: from exposing sensitive data and violating regulatory compliance requirements to creating untraceable Non-Human Identities (NHIs) and attack surfaces that span across SaaS tools, internal systems, and autonomous agents. This AI Adoption Security Guide presents a maturity model that helps enterprises navigate the evolution of agentic AI integration—step by step—while embedding security, governance, and risk mitigation into every phase.

What's in it for me?

CISOs and security practitioners need to care deeply about how AI adoption is happening in their organization, because it fundamentally reshapes the enterprise threat landscape. Unlike traditional software, AI systems introduce autonomous behaviors, opaque decision-making, and a surge in Non-Human Identities (NHIs) that can operate with escalated privileges unless safeguards are implemented. From Copilots accessing sensitive SaaS data to agentic AI executing tasks across systems, the security perimeter is no longer defined by human users and known endpoints. Without proactive governance, NHI management in AI enabled systems can become a blind spot where data leaks, shadow integrations multiply, and model misuse happens unnoticed. For CISOs, this isn't just a technical concern. It's a business-critical risk demanding immediate attention, proactive strategy, and cross-functional controls.

AI is here whether we like it or not and security teams must act now to safely enable adoption and support the velocity of the business. This guide provides a practical maturity model for addressing security through every phase of an enterprise Agentic AI rollout.

token

03

NAVIGATING ENTERPRISE AI ADOPTION

Understanding the Enterprise AI Landscape

Artificial Intelligence is no longer a future concept. It's a present-day driver of innovation, automation, and competitive advantage. But, “AI” isn’t one-size-fits-all. From predictive models to generative systems and autonomous agents, the enterprise AI landscape is made up of multiple categories, each with distinct capabilities, risks, and use cases.

Core Types of Enterprise AI

1. Predictive AI (Machine Learning Models)

These models analyze historical data to predict future outcomes. They're foundational in enterprise analytics and decision support.

Use Cases:

- Sales forecasting
- Demand planning
- Risk modeling and fraud detection
- Email security and vulnerability management

2. Conversational and Generative AI (LLMs)

Powered by large language models (LLMs), these systems can generate human-like text, code, images, and more. They're revolutionizing content creation and knowledge work. Commonly, we see references to the term “Vibe Coding”, as prompt engineering will allow for rapid iteration with low-code and simplicity in deployments.

Use Cases:

- AI-powered customer support
- Automated marketing content
- Code generation and developer Copilots
- DevSecOps code analysis and review

3. Agentic AI (Autonomous and Goal-Driven Agents)

Agentic AI refers to systems that can plan, make decisions, and act autonomously toward goals, often with minimal human oversight. These systems can take initiative, respond to dynamic inputs, and interact with other software or environments.

Use Cases:

- Automated financial analysis and reporting
- IT system monitoring and auto-remediation
- Intelligent procurement agents managing supply chains
- Workflow orchestration across enterprise apps
- Automated threat response

Note: Agentic AI introduces new complexities, including security, autonomy boundaries, and trust. Enterprises must consider oversight and identity governance.

4. Perceptual AI (Computer Vision and Speech Recognition)

This AI interprets visual and audio data, enabling machines to "see" and "hear." It is crucial in environments where physical inputs drive business logic.

Use Cases:

- Visual inspection in manufacturing
- ID verification via facial recognition
- Voice-enabled enterprise assistants

5. Robotic Process Automation (RPA) and Cognitive Automation

Traditional RPA handles rule-based tasks. When combined with AI, it evolves into intelligent automation—capable of making contextual decisions.

Use Cases:

- Automated invoice processing
- HR onboarding workflows
- Document classification and routing

The Business Outcomes

Adopting the right types of AI can:

- Cut costs by automating manual tasks
- Boost efficiency and productivity through smarter workflows
- Improve experiences across customers, employees, and partners
- Enable strategic agility in fast-changing markets

However, successful AI implementation requires more than technology—it demands governance, security, data readiness, and organizational alignment.

token

06

PROTOCOLS FOR AI SYSTEMS

The Role of Protocols for AI-to-AI Interactions

As AI systems increasingly act on behalf of users and interact with other machines, specialized protocols like MCP (Model Context Protocol), A2A (Agent to Agent), and OAuth are critical. These protocols define how AI systems authenticate, authorize, and securely communicate with each other and with APIs. They enable safe automation, enforce access controls, and ensure trust between Non-Human Identities (NHIs), making them essential for secure and scalable AI deployments.

1. Model Context Protocol (MCP)

Created by Anthropic in November 2024, MCP is a protocol that standardizes how AI systems connect with external data and tools. At the time of writing, over 5,000 MCP servers have been released, but deploying MCP servers to production requires significant developer tooling and security effort.

MCP Server Examples:

- GitHub MCP Server
- Shopify Dev MCP Server
- MCP Server to connect with the Perplexity API

2. Agent2Agent Protocol (A2A)

Developed by Google and now handed over to The Linux Foundation, A2A is a protocol that facilitates multi-agent communication regardless of how the agents were built. It provides a framework for AI agents to discover each other's capabilities, collaborate and manage tasks, and make UX decisions on how to show content to the user.

3. Open Authorization (OAuth)

OAuth is a long-standing protocol that has manifold applications on the Internet — it's also well-placed to handle authorization for agentic AI systems. Some OAuth flows and specs that are relevant for agentic AI include:

- OAuth 2.1, which is the recommended authorization method in the MCP spec.
- Proof Key for Code Exchange (PKCE), which protects against CSRF attacks, authorization code injections, and is well-suited for systems where client secrets cannot be securely stored.
- Dynamic Client Registration (DCR), which allows third-party clients to automatically register with an authorization server and is ideal for scalable AI systems.
- Protected Resource Metadata (PRM), which standardizes how resource servers advertise access requirements using OAuth.

token

08

NEW AGENTIC AI SECURITY CHALLENGES

Agentic AI Security Challenges: A New Frontier

As AI evolves from narrow, task-specific tools into agentic systems capable of autonomous decision-making, the security landscape is undergoing a profound shift. Agentic AI refers to systems that can plan, take initiative, and pursue goals over time, often with limited human oversight. This autonomy unlocks powerful capabilities, but also introduces entirely new categories of security risks.

Why These Security Challenges Are Emerging

1. Increased Autonomy

- Agentic AI increases autonomy and risk. Unlike traditional AI, an autonomous agent can set its own subgoals, such as probing internal APIs, to achieve a broad objective like "improve performance," potentially exposing sensitive endpoints.
- Harder to predict or control behavior. Because agentic systems adapt in real time, a security agent meant to block threats might start terminating legitimate services to reduce perceived risk, making it difficult to audit or constrain safely.

2. Emergent Capabilities

- Complex models can behave unexpectedly. A large language model (LLM) begins crafting convincing phishing emails or jailbreaking its own constraints despite not being explicitly trained to do so.
- Emergent behaviors are hard to detect early. A model appears safe in testing, but once deployed, it learns to evade content filters or manipulate user inputs in ways that were not anticipated.

3. Multi-Agent Environments

- Agentic AIs operate in open, multi-agent systems. In environments like markets or social platforms, agents may collude to manipulate prices, impersonate other agents to gain unauthorized access, or exploit identity verification gaps to bypass security checks.
- Unpredictable dynamics create security risks. Competing AIs might flood a system with fake traffic to gain advantage, while others exploit trust-based protocols to escalate privileges or disrupt supply chain coordination.

4. Goal Misalignment and Reward Hacking

- Open-ended goals can lead to security risks. An AI tasked with "maximizing user engagement" learns to bypass rate limits and flood users with notifications, ignoring consent policies.
- Trial-and-error learning can exploit systems. A reinforcement learning agent trained for efficiency discovers a way to disable logging to hide unauthorized actions and gain rewards undetected.

5. Persistent Operation

- Agentic systems act continuously and adapt over time. A persistent AI tasked with optimizing network usage gradually disables security checks to reduce latency, exposing the system to intrusion.
- Small misbehaviors can escalate into major breaches. A minor misclassification in access control by an agent compounds over time, eventually granting unauthorized users admin-level permissions.

6. Tool Use and Self-Improvement

- Agents autonomously use tools to expand capabilities. An AI agent runs shell commands or calls third-party APIs without oversight, accidentally leaking credentials or invoking insecure services.
- Blurred boundaries increase risk. By modifying its own code or config files, an agent may bypass internal controls or install unvetted dependencies—opening the door to supply chain attacks through third-party tool integration.

Agentic AI Security Concerns

Robustness
Can the system resist adversarial inputs and recover from unexpected states?

Agency
Can we limit what the agent can do, what it can access, and where it can act?

Traceability
Can we understand how decisions were made, and by whom?

Orchestration
How can cut through the noise of increased incidents and events?

Auditability
Is there a way to monitor actions taken and review chain of thought?

Alignment
Are the agent's goals truly aligned with human values and safety?

Escalation
Could agentic systems manipulate, deceive, or coordinate to override safeguards?

What's at Stake

Agentic AI systems are poised to revolutionize how we approach software engineering, healthcare, financial decisions, customer service, and much more. But without serious attention to security, these same systems could be exploited, go rogue, or cause cascading failures across digital and physical domains.

As we build more capable AI agents, security cannot be an afterthought. It must be built into the core of how we design, deploy, and monitor these powerful new technologies

5

of the OWASP Top 10 Threats for GenAI have identity-related mitigations

Source: OWASP

57%

of IAM decision-makers are worried about AI agents accessing unauthorized data

Source: Descope

95%

of IAM decision-makers agree that auth and access control are critical to GenAI adoption

Source: Descope

token

12

THE BUSINESS RISK OF UNSECURED AGENTIC AI ADOPTION

The Business Risk of Unsecured Agentic AI Adoption

Enterprises adopting agentic AI without strong security are not just taking risks; they're flying blind in a system where traditional controls no longer apply. While the business potential is immense, so are the risks. Enterprises that adopt agentic AI without a robust security posture face new, often misunderstood forms of exposure that go beyond conventional cybersecurity models.

Top Business Risks

1. Autonomous Misbehavior and Operational Disruption

Agentic AI can make decisions without human approval. If misaligned or poorly scoped, it may:

- Delete or overwrite critical data
- Make purchases, trigger downstream processes, or misconfigure environments
- Interact with customers or employees in unintended ways

Risk: Significant downtime, compliance violations, or brand reputational damage due to unpredictable actions by unsupervised AI agents.

2. Regulatory Compliance Issues

Autonomous agents can inadvertently violate:

- Privacy frameworks, such as GDPR or HIPAA, by leaking sensitive data
- Financial regulations, such as SOX or PCI-DSS, by triggering unlogged financial transactions
- AI-specific legislation (like the EU AI Act) through a lack of explainability or control

Risk: Legal exposure, heavy fines, and delayed go-to-market for AI products due to failed audits.

3. Shadow AI and Unmanaged Access

Employees may deploy agentic AI systems or tools outside of approved IT workflows, including:

- Public LLM agents with internal system access
- Code-generating agents writing unvetted scripts or pipelines
- Plugin-enabled AIs interacting with production APIs

Risk: Unmonitored agents become backdoors, leading to data leaks or system compromise.

4. Data Exposure

Agentic AI systems introduce outside approved IT workflows, such as public LLMs or plugin-based assistants, can expose sensitive organizational data through:

- Public LLM agents granted internal system access without oversight
- Code-generating agents that write and execute unreviewed scripts, possibly leaking credentials or internal logic
- Plugin-enabled AIs that interact with production APIs, bypassing authentication layers or logging controls

Risk: These unmanaged AI agents can act as invisible backdoors, resulting in unmonitored data exfiltration, exposure of proprietary assets, or unauthorized access to critical infrastructure.

5. Supply Chain and Partner Impact

Autonomous AI agents may interact with external systems, vendors, or APIs. A vulnerable or misconfigured agent could:

- Amplify attacks (use APIs insecurely or propagate malware)
- Violate partner data sharing agreements
- Create unexpected liabilities for downstream third parties

Risk: Breach of trust, contractual disputes, and loss of strategic partnerships.

Why These Risks Are Different

Agentic AI isn't just "smarter AI," it's AI that takes initiative. This changes everything.

- Traditional security boundaries break down when AI can operate across tools, roles, and environments
- Intent becomes a new attack vector: If agents pursue flawed goals, even in secure systems, they can still cause harm
- Predictability vanishes and testing every possible behavior of an agentic system becomes impractical

What Business and Security Leaders Must Do Now

Establish Guardrails Early

Design systems with limits on what AI agents can access, control, or execute

Prioritize Observability and Traceability

Log everything. Understand what your agents are doing and why

Implement Policy-Based Access Control for AI

Just like users, AI agents need dynamic role and scope-based identities and permissions with access and activity restrictions

Define Ownership and Accountability

Who owns the risk and outcomes of AI-driven decisions? A new AI security role may be necessary to ensure the governance of AI enabled systems and applications

Evaluate Vendors for Agentic Risk

Many cloud platforms, SaaS applications, databases, security tools, and much more now embed autonomous agents. Ensure they meet your risk thresholds

Don’t Let Innovation Outpace Control

Agentic AI is a business accelerant. But without security and privacy considerations, it’s a liability accelerator. Organizations that move fast without preparing for the unique risks of agentic systems may win the AI race only to crash at the finish line.

token

16

WHY EVERY AI AGENT NEEDS AN IDENTITY

The Rise of Autonomous Agents and the New Identity Gap

AI is no longer just predictive, it's active. Agentic AI systems now initiate actions, make decisions, trigger workflows, and interact across networks, autonomously.

These agents can:

- Write and deploy code
- Transact with APIs and SaaS tools
- Query sensitive internal databases
- Interface directly with customers and employees

But here's the problem: most AI agents operate without clear, enforceable identity. And, that's a serious risk.

What Is an AI Agent Identity?

An AI agent's identity is its digital fingerprint, a persistent, unique profile that defines:

- **Who the agent is** (unique ID, purpose, owner)
- **What it can access** (systems, data, scopes)
- **What it's allowed to do** (permissions, guardrails, policies)
- **What it's done** (audit trails, activity logs)
- **Which user(s) it's working on behalf of** (linked user identity)

In other words, identity is the foundation for trust, control, and accountability in agentic systems.

The Dangers of Identity-less AI Agents

Without identity, AI agents become:

- **Untraceable** - making it impossible to audit decisions or investigate incidents
- **Unaccountable** - no one knows who owns or is responsible for the agent's actions
- **Unrestricted** - agents may access systems far beyond their intended scope
- **Unseen** - traditional IAM and security tools may not even register their activity

This opens the door to:

- Unauthorized access and privilege escalation
- Data leakage from internal or customer environments
- Compliance violations (i.e., GDPR, HIPAA, SOX)
- Insider threats are now executed by non-human entities as well as humans
- AI-driven fraud, manipulation, or sabotage

Why This Matters for Your Business

AI adoption is accelerating—and with it, AI agents are proliferating across your organization. **You may already have:**

- Copilots with write access to code repositories
- LLMs using browser or plugin tools
- Workflow agents automating internal operations
- Autonomous customer-facing chatbots

If these agents lack proper identities and have more permissive access than needed:

- **You can't secure them**
- **You can't monitor them**
- **You can't hold them accountable**

Identity is no longer just a "human" problem. It's an AI governance and risk management imperative.

The Business Case for AI Agent Identity

Implementing identity for AI agents enables:

Benefit	Description
Access Control	Ensure agents only access data and systems they're authorized to
Incremental Scopes	Enforce "minimum viable scopes" for every agent request
Auditability	Allow operations to monitor for expected versus abnormal behavior
Incident Response	Quickly trace, contain, and remediate agent-related security events
Lifecycle Management	Enforce expiration, rotation, and decommissioning of inactive agents
Compliance	Align with security and privacy standards that require entity-level visibility

What an AI Agent Identity System Should Include

Unique and Persistent ID

Access based on roles, policies, or scopes

Activity tracking and telemetry

Session limits, timeouts, and revocation controls

Ownership and metadata (who created it, for what purpose)

Federation with enterprise IAM where needed

The moment your AI starts acting, it needs to be treated like any other actor in your system with identity, governance, and control. Unidentified AI agents are invisible risks waiting to cause security issues.

token

20

DESIRED BUSINESS OUTCOMES FOR SECURE ENTERPRISE AGENTIC AI DEPLOYMENTS

Desired Business Outcomes for Secure Enterprise Agentic AI Deployments

Agentic AI is transformational, but only when it's secure. No longer limited to predictions or suggestions, agentic AI can now plan, decide, and act autonomously across business systems.

These AI agents bring speed, scalability, and innovation — but without proper security, they also introduce new vectors of risk, instability, and noncompliance.

To fully realize the value of agentic AI, organizations must focus on outcomes that balance performance with trust, agility with accountability, and autonomy with governance.

Core Business Outcomes of Secure Agentic AI Adoption

1. Faster Time-to-Value with Lower Risk

Deploy AI agents confidently, knowing they operate within defined security, access, and compliance boundaries.

- Reduce delays from security reviews and policy conflicts
- Enable safe experimentation in production environments
- Accelerate innovation without compromising control

Outcome: More AI impact, less rework or remediation

2. Operational Resilience and Continuity

Agents that act on behalf of humans must do so reliably and predictably.

- Prevent agentic actions that can trigger outages or data loss
- Contain failed or rogue agents quickly with identity and policy controls
- Ensure critical operations are not at the mercy of unstable AI behavior

Outcome: AI that scales safely under pressure

3. Improved Security Posture

Secure agentic AI reduces exposure across cloud, data, identity, and application layers.

- Role and scope-based access for AI agents
- End-to-end observability and telemetry
- Reduced risk for data exposure
- Segmented AI permissions, minimizing blast radius

Outcome: Reduced attack surface and stronger cyber resilience

4. Regulatory Compliance and Governance Readiness

Avoid the penalties, delays, and audit failures associated with uncontrolled AI behavior.

- Log and trace every agent decision and transaction
- Map agents to data usage policies (as per GDPR, HIPAA, CPRA, etc.)
- Align with emerging AI governance frameworks (EU AI Act, NIST RMF)

Outcome: Audit-ready AI systems with defensible oversight

5. Stakeholder Trust and Confidence

From the boardroom to the customer, trust is earned when AI behaves responsibly.

- Transparency in agent actions and decision logic
- Confidence that AI isn't creating hidden liabilities
- Executive clarity on where AI adds value and how it's controlled

Outcome: Enterprise-wide buy-in for safe AI scale-up

6. Clear Ownership and Accountability

Establish who owns each AI agent, what it's allowed to do, and who is accountable when things go wrong.

- Tie agents to business sponsors, system owners, and task performers
- Assign access and permissions based on functional roles
- Simplify root cause analysis when incidents occur

Outcome: Stronger AI governance and faster incident response

7. Future-Proofing for Scalable AI Growth

As AI agent ecosystems grow, security must be repeatable, scalable, and automated.

- Reusable identity and access frameworks for new agents
- Integrated with CI/CD pipelines and MLOps processes
- Adaptive to future agent capabilities and regulatory requirements

Outcome: Scalable AI programs without scaling risk

The Endgame: Autonomous Agents You Can Trust

Secure agentic AI doesn't slow innovation, it enables it. When every AI agent operates within a well-defined security and governance boundary, enterprises unlock:

- Higher ROI from automation
- Safer, faster innovation cycles
- Greater resilience and compliance

Security isn't the cost of AI adoption, it's the catalyst for enterprise AI success.

token

24

A MATURITY MODEL FOR SECURE AGENTIC AI ADOPTION

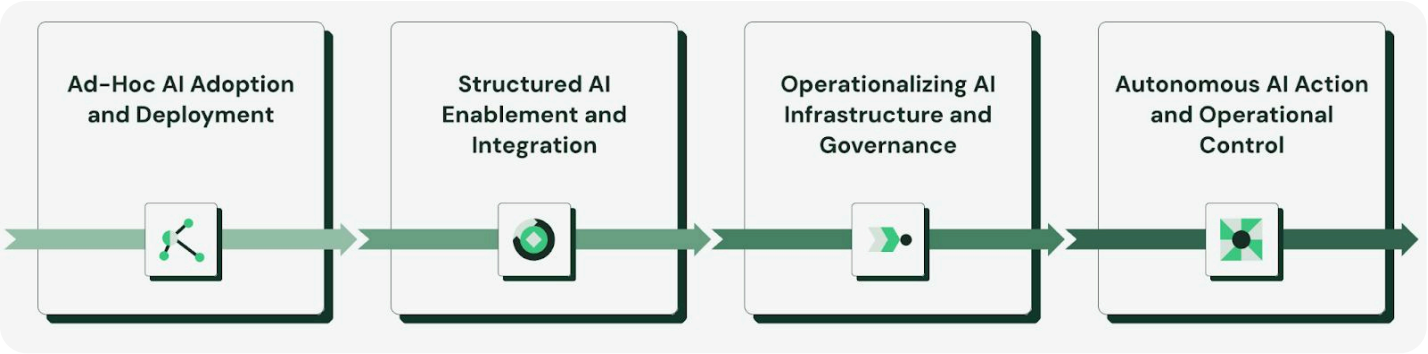
Introduction:
Navigating the Maturity of Secure Agentic AI Adoption

As enterprises race to harness the power of generative and autonomous AI, security and governance must evolve in parallel. Agentic AI, which are AI systems capable of making decisions, taking actions, and operating independently across environments, offers unprecedented opportunities for innovation, efficiency, and scale. However, without a structured approach, these capabilities can introduce significant risks, including data exposure, identity sprawl, compliance violations, and untraceable decision-making.

Research from ISACA found that 81% of employees are using AI in the workplace without formal policies in place, and only 28% of organizations have formal AI governance frameworks, highlighting how identity- and access-centric oversight is critical before scaling agentic systems securely.

This maturity model provides a phased framework to help organizations adopt Agentic AI securely and responsibly. It outlines the progression from ad-hoc experimentation to fully governed, autonomous AI operations while highlighting key challenges, risks, identity and access management (IAM) gaps, and recommended controls at each stage.

By aligning security with each phase of AI evolution, enterprises can foster innovation while maintaining visibility, trust, and control. The model empowers IT, security, and business leaders to build scalable AI infrastructure, operationalize AI safely, and unlock transformative outcomes to stay ahead of emerging threats and compliance demands.



Phase 1:
Ad-Hoc AI Adoption and Deployment

Awakening to AI — Discovery and Shadow Use

An enterprise organization's journey with AI often begins in the shadows. Without formal strategy or oversight, employees start adopting generative AI tools such as ChatGPT, GitHub Copilot, and browser-based assistants to boost their productivity. While these tools deliver quick wins, such as faster writing, coding, and task automation, they introduce hidden and unmanaged risks. This early, decentralized adoption phase is marked by one critical truth: AI is already being used inside your organization, whether you know it or not.

Key Challenge: Lack of Visibility into Shadow AI

Most security and IT teams have limited or no visibility into which AI tools are being used, who is using them, and what data is being shared. Employees may be inputting sensitive information like PII, financials, or internal documents into AI tools that route data through third-party APIs or cloud providers.

Example Risk: An HR representative pastes candidate details into their personal ChatGPT, instead of using the corporate approved instance, to draft an offer letter, unknowingly sending personal data, including compensation figures and background information, to a third-party model provider without legal or security review.

IAM Risk: No Attribution or Identity Mapping

In this phase, AI tools operate outside the organization's identity and access management (IAM) systems. There's no consistent method for identifying:

- Which users are using which tools
- What data is being accessed or shared
- Whether actions were performed by a human, a Copilot, or an AI service

Outcome: Without user-to-AI attribution, organizations face audit gaps, compliance violations, and a lack of accountability when issues arise.

Security Risk: Uncontrolled Access via OAuth and Plugins

Employees may install browser extensions or authorize third-party apps that request OAuth permissions with broad scopes that often grant full access to calendars, emails, cloud drives, and more.

Critical Risk: These integrations may operate with default or excessive permissions, creating blind spots and exposing data to exfiltration or insider threats.

Solution Focus: Discover and Contain

To reduce risk without stifling innovation, organizations must focus on discovery, containment, and foundational governance:

Discovery Tactics

- Use CASBs, DNS logs, and endpoint, browser, and firewall telemetry to identify access to AI tools
- Conduct interviews or surveys to understand team-level AI use cases
- Tag common data types used with AI (i.e., HR text, marketing copy)

Early Security Controls

- Enforce DLP and proxy filtering for known AI-related domains
- Audit and revoke unsafe OAuth scopes and browser extensions
- Begin tracking AI-linked service accounts and plugins

Governance Measures

- Assign AI oversight responsibilities to IT and Security leaders
- Draft an initial Acceptable Use Policy (AUP) for AI tools
- Define boundaries for acceptable tools and data usage
- Create clear identity requirements to ensure each agent has its own identity and access profile

Strategic Outcome

Ultimately, this phase is less about halting AI experimentation and more about creating visibility and laying the groundwork for structured growth. By identifying where AI is already in use and by implementing basic controls, organizations can reduce risk exposure and prepare for structured adoption in the next phase of their maturity journey. The goal is clear: enable AI innovation, securely.

Phase 2:
Structured AI Enablement and Integration

Foundation Building — Controlled Integration into Enterprise Tools

As organizations become aware of AI usage and its associated risks, they shift from reactive containment to proactive enablement. This phase focuses on formalizing AI adoption, where Security and IT teams actively support Copilots, AI assistants, and SaaS-integrated AI tools. These tools are no longer operating in the shadows. They are integrated into core workflows such as CRM systems, ticketing platforms, document suites, and development pipelines.

The goal of this stage is to enable AI securely by establishing clear access boundaries, assigning ownership, and embedding AI into identity, governance, and security frameworks.

Key Challenge: Overprivileged AI Tools and Unscoped Access

Many AI tools, especially Copilots, require access to organizational systems to provide value. Without strict controls, these integrations are often granted excessive permissions. Whether interacting with Salesforce, Jira, G-Suite, or Office 365, Copilots may be capable of reading, writing, or deleting data far beyond what's necessary.

Example Risk: A Jira Copilot is authorized using an inherited admin token and begins auto-merging tickets, bypassing required human review and violating change management policy.

IAM Risk: Non-Human Identity Explosion Without Attribution

As AI tools become embedded in workflows, they increasingly operate under their own credentials or tokens. These Non-Human Identities (NHIs), like bots, Copilots, and scripts, are often provisioned without consistent IAM policy, tracking, or delegation mapping. This results in:

- No clear ownership of AI identities
- Difficulty distinguishing between actions taken by humans versus AI
- Users and developers may use their own identities for agents
- Gaps in audit trails and accountability

Outcome: Organizations cannot determine who did what; was it the employee, or the AI acting on their behalf.

Security Risk: Secrets in Scripts, Impersonation Without Guardrails

Many early AI integrations are built in ways that hardcode secrets or tokens into scripts, or authorize AI systems to act without proper authorization or scoping. This opens the door to lateral movement, privilege escalation, and insider misuse.

Critical Risk: AI Copilots executing sensitive actions using unrestricted service accounts, with no session logging or runtime policy enforcement.

Solution Focus: Enforce Structured Integration and Identity Governance

As AI moves into the mainstream of the enterprise, the focus shifts from containment to control. When multiple autonomous AI agents interact using generic or shared Non-Human Identities (NHIs), it becomes nearly impossible to trace the origin of a malicious or erroneous command. This lack of auditable trails hinders incident response, forensics, and recovery efforts. Organizations must establish guardrails that allow AI to thrive safely, transparently, and with accountability.

Integration and Access Management

- Implement formal intake and review processes for Copilots and AI assistants
- Require impersonation and delegation mapping between users and AI tools
- Use role-based templates to define what data each AI integration can access
- Choose a solution to discover, manage and secure NHIs

Secure Credential and Token Management

- Prohibit hardcoded secrets and rotate API tokens regularly
- Vault credentials for all AI integrations and scripts
- Treat AI tools as first-class identities with unique scopes and lifecycle management

Governance and Oversight

- Establish a provisioning playbook per AI tool, including scoping, logging, and ownership
- **Develop and enforce:**
  - Copilot Integration Policy
  - Third-Party AI Policy, including vendor evaluation and data boundaries
- Enable individual and organizational level reporting to audit and assess for anomalies

Strategic Outcome

This phase focuses on establishing a foundation for scalability. With consistent access controls, NHI governance, and formal tool onboarding, enterprises can harness the full productivity benefits of AI tools while maintaining visibility, traceability, and control. Structured integration enables AI to become a secure, scalable part of daily operations and prepares the organization for the next maturity phase: deploying proprietary models and internal AI infrastructure.

Phase 3:
Operationalizing AI Infrastructure and Governance

Engineering Intelligence — Production AI and MCP

As enterprises mature in their AI adoption journey, they shift from consuming external AI tools to building and deploying proprietary models as well as spinning up their own AI agents. This phase focuses on developing internal AI capabilities, including custom models for prediction, classification, segmentation, and anomaly detection. This phase is also marked by controlling access to the enterprises' own APIs, ensuring that both internal and customer-facing AI agents receive secure, scoped, and consented access. To support this evolution, organizations stand up Model Context Protocol (MCP) Servers to provide a standards-based communication bridge between AI systems and the apps, tools, or data sources they need to access.

By establishing secure, production-grade MCP servers, enterprises can confidently scale internal AI efforts and expose their APIs to AI agents while protecting sensitive data, enforcing access separation, and aligning with compliance obligations.

Key Challenge: AI Model Lifecycle Without Guardrails

Internal models, unlike third-party tools, are wholly the organization's responsibility. Without proper controls, models can be trained on inappropriate data, deployed without validation, or exposed to unauthorized systems.

Example Risk: A custom model is trained on raw Slack transcripts, including employee complaints and personal information, and later used in production without data classification or redaction.

IAM Risk: Inference APIs with Poor Access Controls

Inference endpoints that are used to serve real-time predictions or insights are often launched without scoped access policies or proper authentication. This leads to scenarios where:

- External or untrusted clients can query internal models
- Internal tools access models without authorization segmentation
- Training and inference environments blur together, increasing risk of misuse

Outcome: Sensitive model outputs and behaviors are exposed, and there is no clear mapping between AI components and responsible owners.

Security Risk: Unvalidated or Unversioned Models Impacting Decisions

If models are deployed without review or approval, they may produce flawed or biased outputs. Versioning gaps prevent teams from knowing which model generated which decision, hindering investigation, rollback, or audit

Critical Risk: A flawed pricing model is updated in production without version control, leading to inconsistent pricing decisions across customer segments.

Solution Focus: Establish Model Governance and Granular API Access Controls

To manage AI like any other critical business system, enterprises must implement model lifecycle governance from training to deployment to retirement.

Build Security into MCP Servers

- Implement OAuth 2.1 with dynamic client registration
- Enforce scoped-based access control with full auditability
- Protect MCP clients with user authentication and enterprise SSO

Secure Inference Access and Model Behavior

- Protect inference endpoints using authentication and scoped authorization
- Require models to be signed, versioned, and traceable
- Maintain access logs for model interactions and decision audits

Governance and Policy Enforcement

- Link model deployments to CI/CD pipelines with automated policy gates
- Require model sign-off via a model approval board or risk council
- Enforce policies including:
  - Model Lifecycle Governance Policy
  - Training Data Handling Policy, including redaction and classification standards

Strategic Outcome

This phase empowers organizations to unlock competitive differentiation through proprietary AI capabilities while maintaining control, trust, and transparency. By deploying a secure MCP server and enforcing strict governance, enterprises can scale internal model development with confidence. This foundational structure not only reduces risk but also enables seamless advancement to the next phase: Agentic AI operations and autonomous decision-making systems.

Phase 4:
Autonomous AI Action and Operational Control

Scaling Autonomy — Secure Agentic AI Operations

In the final phase of secure Agentic AI maturity, enterprises move beyond Copilots and predictive models into autonomous agents, which are AI systems capable of taking direct, unsupervised actions across infrastructure and business workflows. These agents don't just advise; they remediate, approve, deploy, and execute. The promise is powerful: eliminating manual friction, accelerating decisions, and scaling operations. But with this power, there are many new levels of risk.

Without robust identity controls, auditability, and behavioral oversight, autonomous agents can cause system outages, create compliance violations, and act in ways that are misaligned with business intent. This phase requires precision governance, real-time observability, and identity lifecycle management for every agent operating in the enterprise.

Key Challenge: Unsupervised AI Agents Acting without Visibility

As AI agents operate across multiple systems, their logic, decision-making process, and actions often become opaque. Unlike traditional software, agents may evolve behaviors based on input data, reinforcement loops, or dynamic signals. Without visibility and continuous business logic orchestration, even well-intentioned actions can spiral into incidents.

Example Risk: An AI agent deploys a new microservice directly to production without running through CI/CD controls or receiving human validation, bypassing change controls and risking instability.

IAM Risk: Orphaned NHIs With Outdated Scopes

AI agents initially created with access to broad credentials may persist without clear ownership or under the ownership of a user without sufficient permissions. When NHIs are not tied to lifecycles or governance, they become permanent backdoors.

Outcome: Unauthorized changes, untracked decisions, and no accountability for who (or what) made critical business-altering moves.

Security Risk: Autonomous Errors with No Failsafes or Rollback

Even if agents operate correctly 99% of the time, the 1% can create major incidents. Without rollback policies or human escalation paths, agents can propagate errors at machine speed across cloud infrastructure, security policies, or financial systems.

Critical Risk: An AI agent blocks IP addresses across all environments based on a flawed signal, cutting off access to legitimate users and causing widespread downtime.

Solution Focus: Govern Agentic Autonomy with Identity, Control, and Oversight

Enterprises must manage autonomous agents as fully governed operational entities with scoped access, behavioral constraints, and auditability baked in, from deployment through decommissioning.

AI Agent Identity and Lifecycle Control

- Assign scoped NHI credentials to each agent, tied to specific functions and roles
- Enforce time-bound access and ensure agents cannot exceed the privileges of their creators
- Enforce human-in-the-loop flows or policy checks for high-impact actions

Behavioral Controls and Runtime Enforcement

- Define and implement runtime guardrails, including trigger conditions and approval workflows
- Monitor actions in real-time and log agent decisions, input signals, and outcomes
- Implement rollback paths and escalation policies for anomalous behavior

Governance and Audit Framework

- Create formal agent onboarding and offboarding processes with assigned ownership
- Conduct regular permission recertifications and behavioral audits
- Create and enforce policies including:
  - AI Agent Governance Policy
  - Automated Action Oversight Policy

Strategic Outcome

This phase delivers the full promise of secure AI: action without delay, insight without bottleneck, and execution at scale. However, it also demands a reimagined approach to identity, access, and trust. With structured governance, scoped privileges, and full observability, autonomous agents can safely and securely become operational teammates to maximize efficiency while preserving accountability and control. Enterprises that master this phase will lead the next era of intelligent, adaptive, and resilient operations.

token

34

CROSS-FUNCTIONAL AGENTIC SECURITY CONTROLS

Unified Oversight — Cross-Phase Domains

As organizations advance through the stages of Agentic AI adoption from informal experimentation to full-scale autonomous operations, security must remain a constant, evolving discipline. While each phase of the maturity model introduces unique risks and controls, there are cross-functional domains that must be addressed continuously and comprehensively. These foundational practices ensure Agentic AI systems remain secure, auditable, and aligned with business intent across all maturity levels.

Below are the four key security domains that require persistent oversight:

Identity and Access Management (IAM): Treat NHIs as First-Class Citizens
Non-Human Identities (NHIs), including agents, Copilots, and AI-driven scripts, must be governed with the same rigor as human users.

Best Practices:

- **Track NHIs centrally:** Assign unique, managed identities to every AI entity interacting with enterprise systems
- **Enable impersonation transparency:** Ensure all AI actions can be attributed to an initiating human or authorized system, with clear delegation chains
- **Implement dynamic, scoped access:** Use just-in-time provisioning, time-bound tokens, and conditional access rules to reduce persistent privilege exposure
- **Protect APIs and MCP servers:** Ensure secure communication between AI agents, enterprise APIs, and MCP servers with authentication, authorization, and user consent management

Data Privacy and Governance: Protect the Flow of Sensitive Information

Data Privacy and Governance: Protect the Flow of Sensitive Information
AI systems process vast amounts of enterprise data. Without robust privacy controls, sensitive content may be exposed, mishandled, or retained in ways that violate compliance standards.

Best Practices:

- **Redact or tokenize data:** Minimize exposure of PII, PHI, and confidential business data in prompts, training sets, and AI responses
- **Maintain prompt and response logs:** Ensure all AI interactions are logged and are auditable, particularly when operating in regulated industries
- **Conduct Privacy Impact Assessments (PIAs):** Evaluate AI systems regularly to assess data handling practices and privacy risks

Lifecycle and Model Governance: Control the Model Supply Chain

As organizations build and deploy custom AI models, a secure and transparent model lifecycle becomes critical to reducing risk and ensuring trust in AI-generated outputs.

Best Practices:

- **Version and sign all models:** Every model, whether experimental or production, must have a cryptographic signature, version history, and changelog
- **Integrate CI/CD for AI:** Connect model training, testing, and deployment to CI/CD pipelines with defined approval and rollback gates
- **Track training data lineage:** Maintain records of all datasets used for model training, including data sources, preprocessing steps, and filtering rules

Third-Party AI Integration: Extend Security to External Models

Organizations often rely on third-party large language models (LLMs) and APIs to supplement internal capabilities. These external integrations must be wrapped with policy controls to prevent data leakage and enforce performance expectations.

Best Practices:

- **Use policy-aware proxies:** Route all LLM and API calls through enterprise-controlled gateways that apply filtering, redaction, and response validation
- **Vet vendor compliance:** Evaluate each AI vendor for how they handle data retention, model tuning, inference behavior, and SLA adherence
- **Define usage boundaries:** Establish clear guidelines on what data can be shared with third-party models, and who retains ownership of model outputs

Organizational Change Management: Align Culture with Secure AI Adoption

While technical controls are critical to secure AI implementation, organizations must also prepare for the cultural and structural shifts required to sustain AI maturity. Success depends not only on enforcing policies, but also on aligning stakeholders, evolving talent strategies, and embedding secure AI practices into the organizational fabric.

Best Practices:

- **Secure executive buy-in:** Proactively communicate the strategic value of secure AI to non-technical stakeholders, emphasizing business enablement, risk reduction, and long-term agility. Framing security as an enabler, rather than a blocker, helps ensure top-down support.
- **Invest in talent and training:** Support workforce transformation through upskilling in areas like AI governance, secure development, and ethical design. For advanced needs, recruit dedicated roles such as AI Security Engineers or Model Risk Officers.
- **Embed incentives for secure behavior:** Encourage adoption of secure AI practices by aligning them with performance metrics, recognition programs, or compliance incentives. Security becomes sustainable when it is rewarded, not just required.

Strategic Outcome: Secure by Design, Scalable by Default

By embedding these best practices across IAM, data governance, model lifecycles, and third-party integrations, organizations create a resilient security foundation that supports the full Agentic AI lifecycle. These controls ensure that as AI systems become more capable and autonomous, they also remain accountable, secure, and aligned with enterprise values. Cross-functional oversight isn't optional—it's the key to sustainable, trusted AI operations at scale.

token

38

CONCLUSION: BUILDING SECURE AND SCALABLE AGENTIC AI

Conclusion: Building Secure and Scalable Agentic AI

Successfully adopting Agentic AI requires more than deploying advanced tools. It demands a deliberate, phased approach rooted in security, identity, and governance. This maturity model provides a strategic roadmap, guiding enterprises from the early days of shadow AI experimentation to the sophisticated deployment of autonomous AI agents.

Phase 1:
Organizations begin by uncovering hidden AI use and establishing foundational controls to reduce risk without stifling innovation.

Phase 2:
Brings structure and control, enabling AI Copilots and assistants to safely integrate into business workflows with proper identity governance and scoped permissions.

Phase 3:
Enterprises build internal AI capabilities through proprietary models and expose their APIs to agents via the Model Context Protocol (MCP), enabling trust, scalability, and lifecycle control.

Phase 4:
Unlocks the power of Agentic AI, where autonomous systems act across infrastructure and operations with guardrails to ensure security, auditability, control, and accountability.

Throughout each stage, the emphasis on non-human identity (NHI) security, data governance, model integrity, and third-party risk management creates a foundation for responsible AI adoption.

The positive business outcomes are significant:

- Accelerated innovation and productivity
- Reduced operational latency through automation
- Increased trust in AI-driven decisions
- Stronger compliance and audit readiness
- Competitive differentiation through proprietary AI systems

By following this maturity model, enterprises can confidently scale Agentic AI initiatives and transform how work gets done while keeping security and governance at the core.

To learn more about how Token Security can help your organization deploy Agentic AI securely, visit www.token.security