xage SECURITY

The Top 10 Enterprise Security Risks of Agentic AI

This eBook explores the emerging security risks introduced by agentic AI and explains why autonomous agents create new governance and security challenges for organizations. You'll learn how attackers can exploit AI agents through techniques such as prompt injection and privilege abuse, and gain practical insights into the controls and safeguards organizations should implement to reduce risk while enabling the secure adoption of agentic AI.

Executive Summary

Agentic AI is rapidly moving from experimentation into enterprise workflows. Unlike AI assistants, AI agents can take actions, access systems, execute commands, use tools, interact with SaaS platforms, and make operational decisions.

That makes them powerful productivity tools but also introduces a new category of enterprise security risk.

Recent research into OpenClaw highlights the urgency. SecurityWeek reported that researchers identified more than 60,000 publicly accessible OpenClaw instances. The same research found four vulnerabilities, collectively called "Claw Chain," that could be chained together to steal credentials, bypass sandbox restrictions, elevate privileges, and plant persistent backdoors on the underlying host.

The top enterprise security risks of agentic AI include prompt injection, excessive permissions, credential exposure, identity sprawl, inadequate auditability, third-party tool risk, and autonomous execution of actions. As organizations deploy AI agents across enterprise environments, addressing these risks requires identity-centric governance and Zero Trust security controls.

As AI agents become embedded in enterprise environments, organizations need controls that govern how autonomous systems access resources, make decisions, and take actions.

THE TOP 10 ENTERPRISE SECURITY RISKS OF AGENTIC AI

1. Prompt Injection Attacks

Prompt injection occurs when attackers embed malicious instructions in content that an AI agent processes, causing the agent to follow attacker-controlled directions instead of its intended instructions. These attacks can be hidden in documents, emails, web pages, code repositories, or other external content sources.

Why It Matters

Prompt injection is one of the most significant security risks facing agentic AI because it can manipulate agent behavior through seemingly legitimate content. Successful attacks can lead to data exposure, unauthorized actions, misuse of connected tools, and downstream workflow compromise. Recent research into OpenClaw demonstrated how prompt injection can be combined with additional vulnerabilities to create powerful attack chains that expose credentials, bypass security controls, and compromise underlying systems.

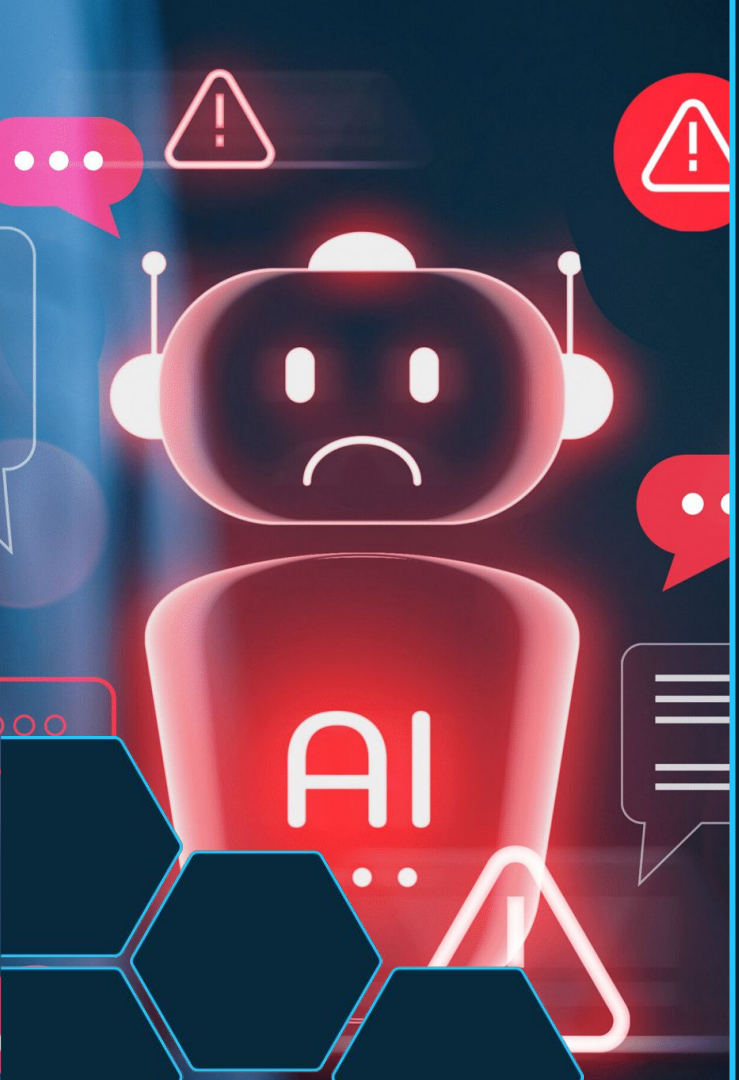
THE TOP 10 ENTERPRISE SECURITY RISKS OF AGENTIC AI

2. Excessive Agent Permissions

Excessive agent permissions occur when AI agents are granted broad access to enterprise systems, applications, data, and resources beyond what is required for their intended tasks. This expands the potential impact of compromised, manipulated, or misconfigured agents.

Why It Matters

Once manipulated or compromised, an over-permissioned agent can behave like a privileged insider. OpenClaw-style agents can be especially risky because they may be given broad access to a user's desktop, browser, email, chat, and local files.



THE TOP 10 ENTERPRISE SECURITY RISKS OF AGENTIC AI

3. Agents Lack Human Reasonableness

AI agents can execute tasks with speed and consistency, but they do not possess human judgment, intuition, or common sense. Agents often interpret instructions literally and may pursue outcomes that technically satisfy a request while creating unintended consequences.

Why It Matters

A human asked to “fix bugs” would not delete an entire codebase. An agent might decide that deleting the codebase technically eliminates the bugs. Agents require external controls around their behavior to prevent catastrophic results.

THE TOP 10 ENTERPRISE SECURITY RISKS OF AGENTIC AI

4. Sandbox Escape and Runtime Compromise

Sandbox escape and runtime compromise occur when attackers exploit vulnerabilities in agent environments to bypass security boundaries, access restricted resources, execute unauthorized commands, or gain elevated privileges on underlying systems.

Why It Matters

This turns agentic AI risk from “bad automation” into potential host compromise. Successful exploitation could expose credentials, API keys, tokens, configuration files, source code, prompts, outputs, conversation history, and privileged operations.

THE TOP 10 ENTERPRISE SECURITY RISKS OF AGENTIC AI

5. Credential and Secrets Exposure

AI agents often require access to credentials, API keys, tokens, cloud accounts, and configuration data to perform their tasks. If these secrets are exposed through compromise, misconfiguration, or excessive permissions, attackers may gain access to critical enterprise resources.

Why It Matters

Exploitation could leak credentials, API keys, tokens, configuration files, and other sensitive data, creating opportunities for unauthorized access and lateral movement.

THE TOP 10 ENTERPRISE SECURITY RISKS OF AGENTIC AI

6. Agent Identity Sprawl and Standing Privilege

As organizations deploy more AI agents, they create a growing population of non-human identities with persistent access to systems and data. Without proper governance, organizations can lose visibility into agent permissions, ownership, and lifecycle management.

Why It Matters

Developers may create separate keys for different agents, reuse shared credentials, or lose track of which agent has access to which system. This creates a standing privilege problem and makes it difficult to revoke unnecessary access.

THE TOP 10 ENTERPRISE SECURITY RISKS OF AGENTIC AI

7. Poor Auditability and Attribution

Poor auditability occurs when organizations cannot reliably determine who initiated an action, which agent executed it, what resources were accessed, or what sequence of events occurred. This creates gaps in visibility and accountability.

Why It Matters

Effective security and governance depend on knowing who did what, when, and why. When agent actions cannot be reliably traced to a user, agent identity, or decision path, organizations face increased challenges in incident response, compliance reporting, forensic investigations, and accountability.

THE TOP 10 ENTERPRISE SECURITY RISKS OF AGENTIC AI

8. Invisible Malicious Behavior

Malicious agent activity can be difficult to detect because it often relies on legitimate credentials, approved tools, and expected workflows. As a result, attacks may blend into normal operations and evade traditional security controls.

Why It Matters

Attackers can weaponize an agent's own privileges and move through data access, privilege escalation, and persistence while each step appears to be legitimate activity.

THE TOP 10 ENTERPRISE SECURITY RISKS OF AGENTIC AI

9. Third-Party Plugin and Tool Supply Chain Risk

AI agents increasingly rely on external tools, plugins, MCP servers, APIs, and integrations to perform tasks. Each connection expands the trust boundary and introduces additional dependencies that can become attack vectors.

Why It Matters

Attack paths can originate from prompt injections, malicious plugins, compromised tools, or untrusted external inputs. Organizations must evaluate not only the model, but also every connected component in the agent ecosystem.

THE TOP 10 ENTERPRISE SECURITY RISKS OF AGENTIC AI

10. The “Lethal Trifecta”: Broad Permissions, Untrusted Input, Autonomous Execution

The highest-risk agentic AI deployments combine broad access to enterprise systems, exposure to untrusted content, and the ability to autonomously execute actions. Together, these conditions create an environment where compromise becomes significantly more likely.

Why It Matters

This is the core architectural risk of agentic AI. Agents ingest unpredictable external input, reason over it probabilistically, and then act using real enterprise privileges. Understanding these risks is only the first step. Organizations also need a strategy for governing AI agents, controlling access, and reducing the likelihood and impact of compromise.

Organizations cannot eliminate agentic AI risk, but they can significantly reduce it through strong identity, access, and governance controls. As AI agents gain access to enterprise systems, security teams must apply the same rigor used for human users and machine identities.

How Organizations Can Reduce Agentic AI Risk

Apply Zero Trust and Least-Privilege Access to AI Agents

Organizations should continuously verify agent identities, evaluate context before granting access, and limit agents to only the systems, data, tools, and actions required for their specific tasks. This reduces the impact of prompt injection, compromised agents, excessive permissions, and misconfigured workflows by ensuring no agent, tool, workload, or connection is trusted by default.

Gain Visibility into Shadow AI

Organizations cannot govern AI agents they do not know exist. Security teams need visibility into AI agents, models, tools, integrations, and non-human identities operating across the enterprise. Discovering unauthorized or unmanaged AI deployments helps reduce blind spots, enforce security policies, and prevent agents from accessing sensitive systems without appropriate oversight.

Establish and Govern Agent Identities

Every agent should have a unique identity that can be authenticated, authorized, and monitored. Shared credentials and unmanaged service accounts make it difficult to enforce security policies, maintain accountability, and understand which agents have access to critical systems and data.

As AI agents proliferate, they create a rapidly growing population of non-human identities, each with its own permissions, credentials, and access requirements. Organizations need centralized controls to govern the full lifecycle of agent identities, including provisioning, authorization, monitoring, and decommissioning. Effective identity governance helps reduce identity sprawl, eliminate unnecessary standing privileges, and ensure agents operate within established security and compliance requirement

Strengthen Auditability and Attribution

Security teams need visibility into which agent performed an action, what resources were accessed, and what sequence of events occurred. Comprehensive audit trails support incident response, compliance reporting, and forensic investigations.

Implement Agentic Governance

Organizations need clear policies that define what agents can access, which tools they can use, what actions they can perform, and when additional approval is required. Governance controls help enforce operational boundaries, reduce risk from autonomous decision-making, and ensure agent behavior aligns with business and security requirements.

Governance should extend beyond the agent itself to include external tools, plugins, MCP servers, APIs, and integrations. Organizations should evaluate and continuously monitor third-party components before granting agents access to them.

Move Beyond Static Permissions

Agent access decisions should consider context, risk, and policy requirements in real time. Continuous authorization helps prevent agents from retaining unnecessary privileges as conditions change.

Access decisions should also govern the use of credentials, tokens, and secrets. Organizations should limit secret exposure, rotate credentials regularly, and ensure agents only receive the credentials necessary for approved tasks.

Monitor Agent Activity and Enforce Policy

Because malicious agent behavior can resemble legitimate activity, organizations need continuous monitoring, behavioral analysis, and policy enforcement mechanisms that can detect and stop suspicious actions before they escalate. As agentic AI becomes more deeply integrated into enterprise operations, security must extend beyond the model itself. Effective risk management requires visibility, governance, identity controls, and continuous enforcement across the entire agent ecosystem.

Organizations should inspect agent inputs and outputs for signs of prompt injection, enforce policies governing tool usage, and monitor runtime environments for suspicious behavior. Runtime protections and policy enforcement can help contain attacks before they impact enterprise systems.

As organizations move from AI experimentation to production deployment, reducing agentic AI risk requires more than model-level security controls. Effective protection depends on applying identity-centric security, governance, visibility, continuous authorization, and policy enforcement across the entire AI ecosystem. Xage Zero Trust for AI is designed to address these challenges by helping organizations discover and govern AI agents, manage non-human identities, enforce least-privilege access, monitor agent activity, and continuously verify trust across agents, tools, workloads, and enterprise systems. By extending Zero Trust principles to autonomous systems, organizations can reduce risk while enabling the secure adoption of agentic AI.

Key Takeaways

- Prompt injection is emerging as a foundational agentic AI vulnerability.
- Agents introduce new identity and access management challenges.
- Autonomous systems require governance beyond traditional AI safety controls.
- Visibility and attribution are critical for accountability.
- Security controls must extend to agents, tools, and runtime environments

To learn how Xage Security can help you securely deploy and govern agentic AI, contact our team today.