



# Redefining security for the agentic AI era

by Kris Lovejoy

Global Security and Resiliency Practice Leader, Kyndryl



# Introduction

The rise of agentic AI marks a monumental leap in technological capability and a fundamental challenge to cybersecurity paradigms.

Autonomous agents can reason, plan and execute complex tasks, enabling enterprises to tackle difficult problems, improve customer experiences and continuously optimize operations. However, the autonomy and adaptability that make these systems so powerful also introduce a new class of vulnerabilities beyond the reach of traditional security models.

For decades, “Secure by Design” has helped enterprises enable resilience by embedding security early. Yet the autonomous nature of agents is now revealing limitations these principles never anticipated.

Today’s leaders face an inflection point: They must evolve their strategies beyond legacy defenses and embrace a new blueprint built for autonomous intelligence. Their ability to realize the potential of agentic AI and mitigate its risks will depend on redefining security for the agentic era.

## Why traditional security models fall short

Secure by Design principles were traditionally predicated on predictable, rule-based systems. They led enterprises to focus on hardening defined perimeters, validating known inputs and preventing exploits of specific code vulnerabilities.

Agentic AI shatters these assumptions. Its dynamic, adaptive and often opaque nature creates a fundamentally different attack surface that renders static defenses inadequate.

While traditional application security focuses on identifying code vulnerabilities, the most severe attacks on AI systems often target training data to corrupt outputs. Traditional security models are also designed for predictable inputs, meaning they’re ill-equipped to defend against adversarial prompts that use cleverly crafted natural language to manipulate agents and override or “jailbreak” safety restrictions.

Agentic AI further skirts traditional approaches by blurring trusted boundaries. Many advanced AI models act with unprecedented speed and agility, making it more difficult to detect when a system has been compromised or is following malicious instructions.

Treating security as a final checkpoint will fall short of securing agentic AI systems that operate across complex technology systems. Rather, agentic AI requires embracing a DevSecOps approach that integrates security throughout the entire development lifecycle, from model training to deployment. Legacy approval processes cannot accommodate the automated, continuous security validation that agentic systems require.

## The growing risk of inaction

The potential consequences of failing to evolve security approaches are severe and multifaceted – inaction is not an option.

Risks now extend beyond traditional data breaches to the manipulation of autonomous systems that can interact with the physical world. An agent operating with broad permissions can be hijacked through subtle prompt manipulation, turning a helpful assistant into a malicious actor capable of exfiltrating data, executing unauthorized financial transactions or causing physical disruption.

Multiagent systems are also susceptible to chain reactions. A single compromised agent can misdirect other agents, leading to a domino effect of systemic failure, misinformation and unpredictable behavior. Compromised agents can enable malicious goals to rapidly spread across interconnected systems, breaching containment boundaries and amplifying harm.

Data poisoning and model theft present additional risks. Attackers may corrupt an agent’s training data to introduce biases or hidden vulnerabilities. Sophisticated adversaries can also reverse-engineer proprietary models through repeated queries, compromising intellectual property.

The autonomous nature of AI agents also makes traditional compliance frameworks insufficient. Without proper enterprise controls, agentic AI systems that process sensitive data may expose organizations to compliance and regulatory lapses. Violating regulations like the General Data Protection Regulation (GDPR) can result in substantial fines, loss of certifications and reputational damage.

The Open Web Application Security Project (OWASP) Top 10, a list of the [most critical security risks](#) for large language models – which serve as the reasoning engine of agentic AI underscores many of these emerging threats, including prompt injection, training data poisoning, and excessive agency. Given these risks, leaders face an urgent imperative to adopt a new security blueprint.

# A new blueprint for securing agentic AI

Government agencies and industry leaders have begun to formulate new frameworks to help enterprises advance agentic AI. Organizations like the Coalition for Secure AI (CoSAI), the U.S. Cybersecurity and Infrastructure Security Agency (CISA), the National Institute of Standards and Technology (NIST) and major tech companies like Google and Microsoft have all contributed to a growing consensus on what a modern Secure by Design approach must entail.

Securing agentic AI requires a multilayered strategy that extends beyond traditional security to address the unique lifecycle and operational realities of autonomous intelligence. A unified framework for secure agentic AI involves four core pillars: foundational governance, secure development lifecycle, robust operational security and adaptive monitoring and response.

## 01. Foundational governance and oversight

Enterprises must establish “the rules of the game” before an agent begins to play. This is why robust governance anchored in people-centric control and accountability is critical to securely managing agentic AI. To build this strong foundation, organizations should automate low-risk decisions while defining boundaries that immediately trigger human intervention. They must also define clear accountability structures throughout the full agent lifecycle, enforce operational boundaries and data access and maintain immutable<sup>1</sup> logs for transparency and auditability. Systems should be designed with built-in compliance measures<sup>2</sup> – such as retention policies and audit trails – and independent verification to meet regulatory standards, including SOC 2, ISO 27001, ISO 23984 and industry-specific mandates.

---

<sup>1</sup> Implementation focus: Use managed logging services such as AWS Cloudwatch logs/ Kinesis, Azure Monitor, GCP Cloud Logging) and enforce immutable storage policies (S<sup>3</sup> Write Once Read Many (WORM), Azure Blob Storage Immutability) to meet audit requirements efficiently. Mandate deployment of XAI (Explainable AI) services like Vertex AI Model Monitoring or wrapper models to generate post-hoc explanations for high-risk agent decisions. Treat the lack of an auditable explanation as a security violation.

<sup>2</sup> Implementation focus: Adopt dynamic compliance reference architectures with automated control mapping to SOC<sup>2</sup> / ISO 27001, powered by agent-aware CSPM (cloud security posture management) tools. Integrate continuous evidence collection and real-time reporting using tools like AWS Audit Manager, Azure Policy Compliance and GCP Assured Workloads, with agentic policy enforcers mapped to regulated data types and agent actions.

## 02. Secure AI development lifecycle

Organizations must embed security into every phase of AI development and secure the entire AI supply chain. This work enables the integrity and quality of training and testing data, verifying third-party tools and pre-trained models, and proactively testing agents’ resilience against adversarial attacks. “Privacy by Design” builds privacy into every stage of the data lifecycle, and incorporating automated security testing early such as model scans, prompt-injection testing, and bias detection helps enable more sophisticated AI systems that are secure long before they reach production. Additionally, zero-trust credential management can allow for continuous verification of agent access.





### 03. Robust operational and runtime security

Securing agentic AI during active use requires real-time enforcement of governance principles. Organizations should treat all agent inputs as untrusted and validate outputs, and apply zero trust principles to verify agent identity and segment agent workloads into isolated environments. Enforcing governance further requires implementing real-time monitoring and content filtering, establishing fail-safe mechanisms to prevent unintended consequences when anomalous behavior is detected, and enforcing strict access controls for minimum access to data, tools and APIs. Organizations should also use API gateways with rate limits, authentication, and payload checks tailored to AI-generated requests.

### 04. Adaptive monitoring and response

Given that threats will constantly evolve, security must be an ongoing, adaptive process. Organizations must design agents for comprehensive<sup>3</sup> telemetry to enable real-time threat detection, forensic analysis and automated response. Automated defenses, learning from feedback loops to refine security controls, and continuous monitoring for abnormal behavior all strengthen the resilience of agentic AI.

## Conclusion

The age of agentic AI has arrived. Organizations now have a profound responsibility to develop and deploy these powerful systems securely. The Secure by Design principles that have served enterprises in the past are insufficient for this new reality. By embracing a holistic, lifecycle-based approach that prioritizes governance, secure development, robust operational controls, and adaptive monitoring, organizations can realize the immense potential of agentic AI without sacrificing safety or security. The call to action is clear: now is the time to recalibrate security for the agentic AI era.

---

<sup>3</sup> Implementation focus: As a cost management strategy, implement tiered telemetry capture such that high-risk agents log all internal steps and low-risk agents log only final actions and security events. Use sampling /edge processing for non-critical data e.g., Azure Stream Analytics, AWS IoT Greengrass, to reduce cloud ingress/storage costs.



# kyndryl<sup>®</sup>

© Copyright Kyndryl, Inc. 2025

Kyndryl is a trademark or registered trademark of Kyndryl, Inc. in the United States and/or other countries. Other product and service names may be trademarks of Kyndryl, Inc. or other companies.

This document is current as of the initial date of publication and may be changed by Kyndryl at any time without notice. Not all offerings are available in every country in which Kyndryl operates. Kyndryl products and services are warranted according to the terms and conditions of the agreements under which they are provided.