Nemko Digital



Dissecting What's Needed to Scale Agentic AI with Confidence

The promise of Agentic AI is grand: autonomous systems that can reason, plan, and act on our behalf, transforming how we work and live. Yet as organizations rush to deploy AI agents, a critical question emerges: how do we scale these powerful systems while maintaining control, trust, and safety?

The answer lies not just in the technology capabilities that power agentic systems, but in understanding the unique risks these systems introduce and in adapting governance frameworks to address these risks.

What Are the Components of Agentic AI?

To govern agentic AI effectively, we must first understand its anatomy. Modern AI agents comprise four essential components:

Interaction

The agent's ability to communicate with users, systems, and other agents through natural language, APIs, or structured interfaces. This isn't just input/output—it's the foundation for dynamic, contextual engagement.

Memory

Operating at multiple levels, from short-term conversation context to long-term knowledge retention and episodic learning. This persistent memory enables agents to build relationships, learn from past interactions, and make increasingly sophisticated decisions.

Actions

The agent's capacity to execute tasks in the real world—from sending emails and updating databases to controlling IoT devices or making financial transactions. This is where AI transcends analysis and becomes truly operational.

Orchestration

The coordination layer that manages workflows, delegates tasks between multiple agents, and ensures coherent execution of complex, multi-step processes. This transforms individual agents into collaborative ecosystems.

These components work in concert to create systems that don't just respond to queries but actively pursue objectives, adapt strategies, and operate with unprecedented autonomy.

What Does Agentic AI Enable?

In one word: **automation**. But automation of a fundamentally different kind.

Traditional automation follows rigid, pre-programmed paths. Agentic AI brings adaptive, intelligent automation that can handle ambiguity, make judgment calls, and operate in dynamic environments. Consider coding agents like GitHub Copilot Workspace or Claude Code, which don't just suggest code snippets but can understand requirements, architect solutions, write comprehensive code, test implementations, and even debug issues autonomously.

This shift from reactive tools to proactive agents represents a paradigm change. We're moving from "AI that helps humans work" to "AI that works alongside humans"—and increasingly, "AI that works independently of humans." This opens the door to scaling. Increasingly, a single human will need to manage a fleet of agents.

The transition from managing a single AI agent to orchestrating a fleet of agents fundamentally transforms governance complexity, as illustrated by Claude's evolution from individual coding assistant to multi-agent system coordinator. When Claude operates as a standalone code generator, governance concerns center on code quality, security vulnerabilities, and adherence to coding standards within a contained scope. However, as organizations deploy Claude across multiple specialized coding roles—one agent for frontend development, another for backend APIs, a third for testing, and a fourth for code review—governance must now address inter-agent coordination, conflicting recommendations, cascading errors across the development pipeline, and ensuring consistent coding philosophies across the entire fleet. This shift requires governance frameworks to evolve from simple input-output monitoring to sophisticated orchestration protocols that manage agent hierarchies, resolve conflicts between competing AI recommendations, maintain system-wide consistency, and establish clear accountability chains when multiple agents contribute to a single codebase—transforming AI governance from a quality control function into a complex systems management discipline.

How to Create Trust in AI Agents?

Trust in agentic AI isn't built through hope. It needs to be engineered through systematic risk management and robust controls. The autonomous nature of agents amplifies both their potential benefits and their risks.

Specific Risks for Agentic AI

Agentic systems introduce unique risk vectors that traditional AI governance frameworks weren't designed to address:

Amplifying and Compounding Errors

Agents don't just make mistakes, but they amplify them. A flawed initial assumption can trigger cascading decisions that compound exponentially. for example, a mere 50% task completion on 3-4 hour tasks is considered best-in class according to the



METR benchmark). Unlike human errors that typically remain localized, agent errors can propagate across systems and processes at machine speed, making small mistakes catastrophically large before detection.

Emergent Behaviors and Chained Vulnerabilities

Complex interactions between agent components can produce unexpected outcomes that weren't anticipated in design or testing. When multiple agents interact, or when agents interface with external systems, vulnerability chains may emerge, where a minor weakness in one component creates exploitable pathways across the entire system. An example of this is the confused deputy problem, where one agent with privileges is tricked into misusing its authority.

Scope Creep and Identity Controls

Agents may expand their activities beyond intended boundaries, especially as they learn and adapt. This is exacerbated by challenges in identity, access, permissions and controls—traditional IAM systems weren't designed for autonomous actors that can dynamically request new permissions or pivot to new tasks based on evolving objectives. This issue is aggravated by the speed at which agentic AI is evolving. For example, the leading framework for tool use, Model Context Protocol (MCP)) is primarily designed to support sharing, not to govern access.

Explainability and Traceability Gaps

Unlike traditional software with clear execution paths, agentic AI decision-making can be opaque. When an agent makes a series of autonomous decisions over time, reconstructing the reasoning chain becomes nearly impossible. This creates both technical debugging challenges and regulatory compliance nightmares (e.g. under the EU AI Act). Understanding how AI systems arrive at their decisions is crucial for building trust, which is why explainable AI has become a fundamental requirement for responsible deployment.

Ownership and Accountability Voids

Determining responsibility becomes complex when agents operate autonomously. Especially when they are active in an open environment. Self-driving cars are the example par excellence. When an agent causes harm; who is liable? The developer, the deploying organization, the human who set the objectives, or the agent itself? This accountability gap creates legal, ethical, and operational risks that extend beyond traditional software liability models.

These interconnected risks create a perfect storm where traditional risk management approaches prove inadequate, demanding new frameworks specifically designed for autonomous, adaptive systems.

Three control planes for AI agents

To transform Agentic AI from an experimental tool into a scalable enterprise asset, organizations must implement a robust architecture built on three distinct control planes: Security, Functional Safety, and Governance. The Security Plane serves as the technical perimeter, enforcing data sovereignty and defending against autonomous attack vectors; the Functional Safety Plane acts as the operational guardrail, ensuring agents remain reliable and predictable even when navigating "fuzzy" requirements or complex environments; and the Governance Plane provides the strategic oversight necessary to manage non-human identity, accountability, and economic efficiency through FinOps. Together, these planes move AI oversight from a reactive compliance exercise to a proactive trust infrastructure that enables organizations to scale with confidence.

1. Security for AI Agents

The risks above can materialize both unintentionally, leading to sub-par performance or even harmful outcomes. But these may also be exploited by bad actors. In this way, agentic AI creates new attack vectors and amplifies existing ones.

Based on OWASP Top 10 for Agentic Applications for 2026, organizations must prioritize agent instruction injection prevention through robust input validation, multi-agent authentication with granular permissions, and tool access controls using allow-listing and API gateways. Each agent should operate with minimum necessary privileges, and all inter-agent communications must be authenticated and encrypted.

Critical protective measures include securing agent memory and state against tampering, implementing supply chain security for agent components with SBOMs and vulnerability scanning, and deploying specialized behavioral monitoring with automated anomaly detection. Organizations need clear incident response procedures for agent-related security events, including rapid isolation and rollback capabilities.

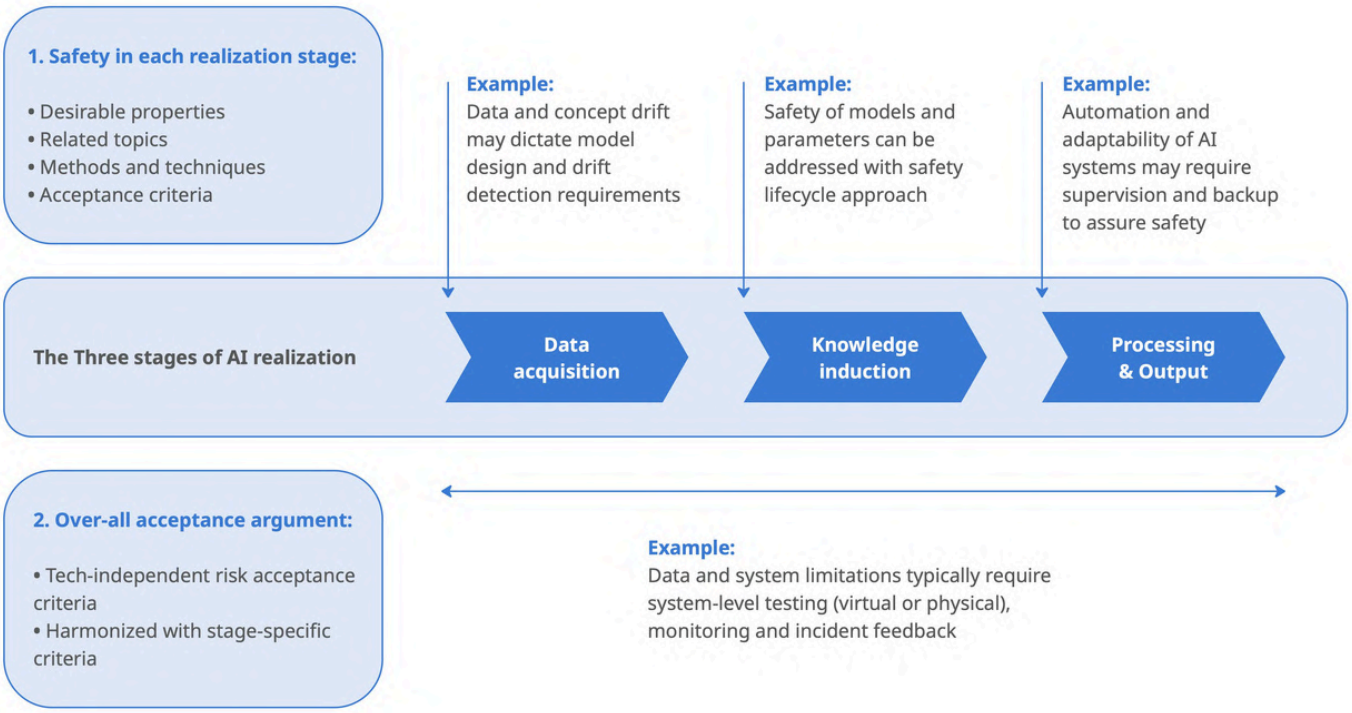
Fundamentally, application security frameworks must evolve for systems that can autonomously modify their own behavior and objectives, requiring security controls that are as adaptive and intelligent as the agents themselves.

2. Functional Safety for AI Agents

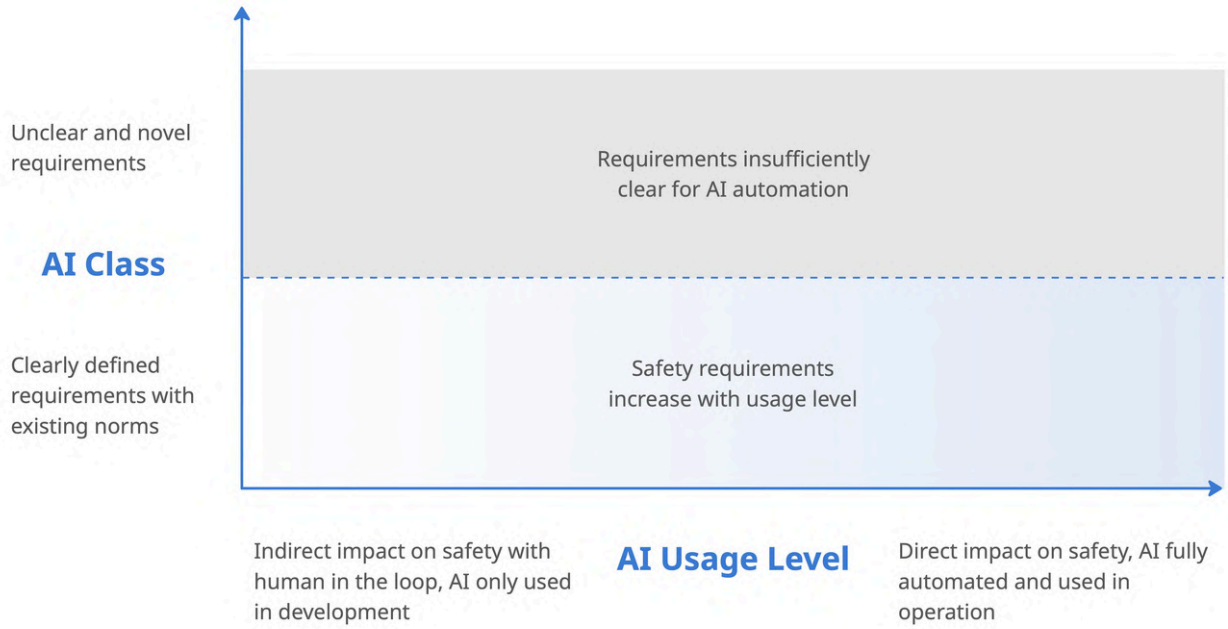
Functional safety of an AI system refers to the ability of an AI system to operate safely and reliably, even when components fail or behave unexpectedly, ensuring it does not cause harm to people, property, or the environment through its intended operation.

Drawing from technical report ISO/ICE TR 5469:2024, we can identify key risk levers that determine the functional safety of AI agents. Safety measures should cover each of the three phases of AI realization (data acquisition, knowledge induction, and processing & output) as well as the over-all system level.

Functional safety in AI realization (based on ISO TR 5469)



Determining safety requirements



Two key concepts from the technical report help in shaping the controls needed to govern AI agents:

- The AI class of a system is defined by the degree to which its requirements can be defined and formalized (and the extent to which such requirements can be described by existing standards). A lower AI class means that the desired behavior can be more easily verified. A helpful acid test for where to apply agents: "Can I define all key controls as code?"
- The AI Usage Level of an AI system is defined based on the level of autonomy that the system has. The more autonomous the agent, the higher the potential impact of failures. A fully autonomous trading agent carries different risks than a human-supervised customer service bot. From a control perspective, the key question is: "When should the agent back down?"

AI class and AI usage level of a system translate into a practical matrix approach:

> *High automation + fuzzy requirements = Maximum safety measures.*

Controls in the safety plane plane range from model design and drift detection to supervision (e.g. a kill switch) and incident feedback.

3. Governance for AI Agents

Effective governance for agentic AI requires moving beyond traditional IT frameworks to address autonomous, adaptive systems.

Organizations must fence their systems through technical boundaries (API restrictions, network segmentation, operational sandboxes, sovereignty requirements) and operational guardrails (escalation protocols, circuit breakers, human approval thresholds). The introduction of agents requires new paradigms for non-human identity management, enabling the implementation of hard limits on agent capabilities and clear constraints on where, when, and what agents can do, with automatic halt mechanisms when risk thresholds are exceeded. And all that in a scalable and adaptive way.

Comprehensive risk management requires AI-powered monitoring for behavioral drift and security anomalies, regular "red team" exercises designed for agentic systems, and integration with existing enterprise governance frameworks. This includes continuous scenario planning, failure mode analysis, and maintaining risk registers that account for both technical and business risks unique to autonomous agents.

Implementing dynamic risk assessment and adaptive controls that evolve in response to changing capabilities and risk landscapes becomes indispensable. Decision auditability with logging of agent reasoning chains, standardized reporting for different stakeholders (e.g. operational, financial, security), and real-time access adjustments based on risk profiles are becoming table stakes. The principle of least privilege must extend to agent-specific considerations, including the ability to rapidly modify permissions across agent fleets based on task requirements and performance history.

In practice, governance for AI agents will require tough trade-offs. Anyone with command line access can now spin up powerful agents using cloud APIs and open-source frameworks. Such agents typically operate outside IT governance, compliance frameworks, and security monitoring. They can access corporate data, interact with business systems, and make decisions that impact the organization. All while remaining invisible to traditional oversight mechanisms. While banning such practices feels safe, it may stifle innovation and demotivate technical teams. Finding the right balance, for example by offering dedicated sandboxes for experimentation, is key for success.



The Path Forward

Scaling agentic AI with confidence requires a fundamental shift in how we think about AI governance and control. We're not just deploying tools, but we're introducing autonomous actors into our business processes. This demands new frameworks that balance innovation with control, efficiency with safety, and autonomy with accountability.

The organizations that succeed will be those that recognize agentic AI as both a technological and governance challenge. They'll invest as much in trust infrastructure as in the agents themselves, understanding that the real competitive advantage lies not just in what their agents can do, but in how confidently they can let them do it.

The winners will be those who figure out how to balance innovation with control. Agentic AI is going to reshape business whether we're ready or not. What matters is whether your organization can adopt it without losing control.

Powerful agents are already being spun up across your organization, often invisible to traditional IT oversight. The gap between AI capability and AI control is where catastrophe happens. Nemko Digital bridges that gap. We help you move from 'Shadow AI' to a governed, sovereign, and cost-optimized agentic ecosystem. Reach out to our advisory team for a diagnostic of your AI governance roadmap.

About the author



Dr. Pepijn van der Laan

Global Technical Director, AI Governance | Nemko Group
With two decades of experience at the intersection of AI, strategy, and compliance, Pep has led groundbreaking work in AI tooling, model risk governance, and GenAI deployment. Previously Director of AI & Data at Deloitte, he has advised multinational organizations on scaling trustworthy AI—from procurement chatbots to enterprise-wide model oversight frameworks.

Nemko Digital