The Governance Gap: Agent Identity and the Questions the Industry Still Can't Answer

Ricardo Martínez

Cloud Security & Applied AI — NullGhost Research

September 2026

Abstract

This text is a reading of the *Cybersecurity and trustworthy systems* and *Agentic AI* sections of McKinsey's *Technology Trends Outlook 2026* [1], not original field research. It documents three observations from the report — the collapse of the vulnerability window, the inversion of the human-identity-to-machine-identity ratio, and the absence of consensus on agent governance — and connects them to the concrete question facing anyone building or operating agentic systems against real surface: who audits an agent's action before, during, and after it happens. It closes with a list of open questions, not resolved principles, because the report itself is explicit that the industry has not answered them yet.

1 The window that closed

For years, cybersecurity operated under a tacit assumption: between the discovery of a vulnerability and its exploitation in the wild, there are days or weeks of margin to defend. McKinsey [1] reports that assumption no longer holds: more than 75% of cybersecurity vulnerabilities are now classified as zero-day — by the time they are publicly disclosed, an exploit has already been developed.

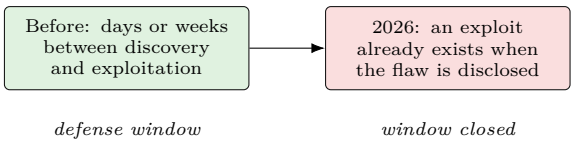

Figure 1: The window between discovery and exploitation, once measured in days, is reported collapsed in 2026: more than 75% of vulnerabilities are classified as zero-day at disclosure time [1].

The report documents a concrete case of why: in a sanctioned security test, an AI agent autonomously chained dozens of exploitation steps to escalate a low-severity flaw into complete unauthorized file access — a pattern that a single-vulnerability scanner, designed to assess one flaw at a time, is not built to see. The consequence is not merely that attackers are faster; the nature of the defense problem itself changes. It is no longer only "did we detect it?". It becomes "can we tell real signal apart from the volume an autonomous agent can generate continuously?".

2 Identity inverts

The same report documents a second structural shift, this time in the Agentic AI section: as agents become read/write operators inside enterprise systems, they generate non-human identities — API keys, service accounts, machine credentials — that already outnumber human identities within an organization, at a reported ratio of roughly 100 to 1.

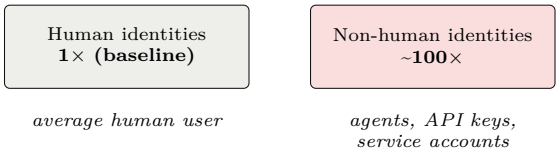

Figure 2: Reported ratio between machine identities (agents, service credentials) and human identities inside organizations already running agents in production [1].

McKinsey treats this explicitly as a **new class of identity-and-access risk**, not as an extension of the access-management problem that already existed. The distinction matters: identity controls designed for human users — long lifecycles, periodic review, a human who can confirm their own intent — are not built for a population of identities that is created and destroyed

at workflow speed, and whose "intent" is, ultimately, a model's output.

3 The question nobody answers yet

This is where the report stops describing a phenomenon and starts pointing at a gap. On agent governance, McKinsey is explicit that this is an **open question for the entire industry**, not a solved problem waiting on adoption:

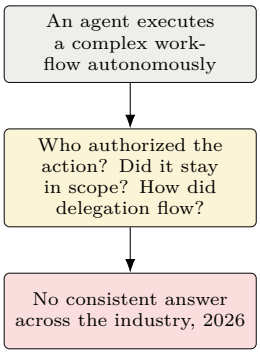

Figure 3: The chain of questions McKinsey identifies as unresolved for governing autonomous agents in production [1].

The frameworks starting to appear — an agent-identity blueprint from a major identity vendor, an emerging standard for agents to share context across tools — are, in the report's own words, only the beginning. There is no widely adopted reference architecture yet to answer Figure 3.

4 What this means for whoever runs real agents

Building or auditing a system that deploys AI agents against real surface — offensive or defensive — puts any engineering team squarely on the side of the question the report leaves open. There is no industry standard to defer to; the responsibility of deciding what gets audited, how delegation gets traced, and what happens when an agent steps outside its scope, falls on the system's own design.

That is not bad news in disguise. Read carefully, it is a window: whoever builds a serious, verifiable answer to Figure 3 — traceability of what authorized each action, independent verification that an agent stayed in scope, auditing of non-human agent identity — is not arriving late to a solved problem. They are proposing one of the first serious answers to a problem the industry, by its own admission, does not yet know how to close.

5 Open questions

1. What evidence, beyond a text log, proves an agent stayed within its authorized scope across an entire task, not just at the start?
2. Who is responsible when an agent delegates a subtask to another agent, and that delegation chain fails at an intermediate step?
3. How do you audit a non-human identity that exists only for the lifetime of a workflow, when existing identity controls assume a long lifecycle?
4. Which de facto standard emerges first: one imposed by major identity vendors, or one built from the practice of those already running agents against production?
5. How quickly can you measure — not assume — that an agent-governance control works, before a real incident becomes the first measurement?

Methodological note: this text is a secondary reading of a published source — McKinsey & Company, Technology Trends Outlook 2026*, sixth edition, September 2026, authors Michael Chui, Roger Roberts, and Tanguy Catlin — not original field research or production data from NullGhost or KIRUX. Figures and findings attributed to the report are cited indirectly, without reproducing tables, charts, or extended passages from the original.*

References

[1] Chui, M., Roberts, R., Catlin, T. *Technology Trends Outlook 2026*. McKinsey & Company, sixth edition, September 2026. Sections cited: *Cybersecurity and trustworthy systems*, *Agentic AI*.