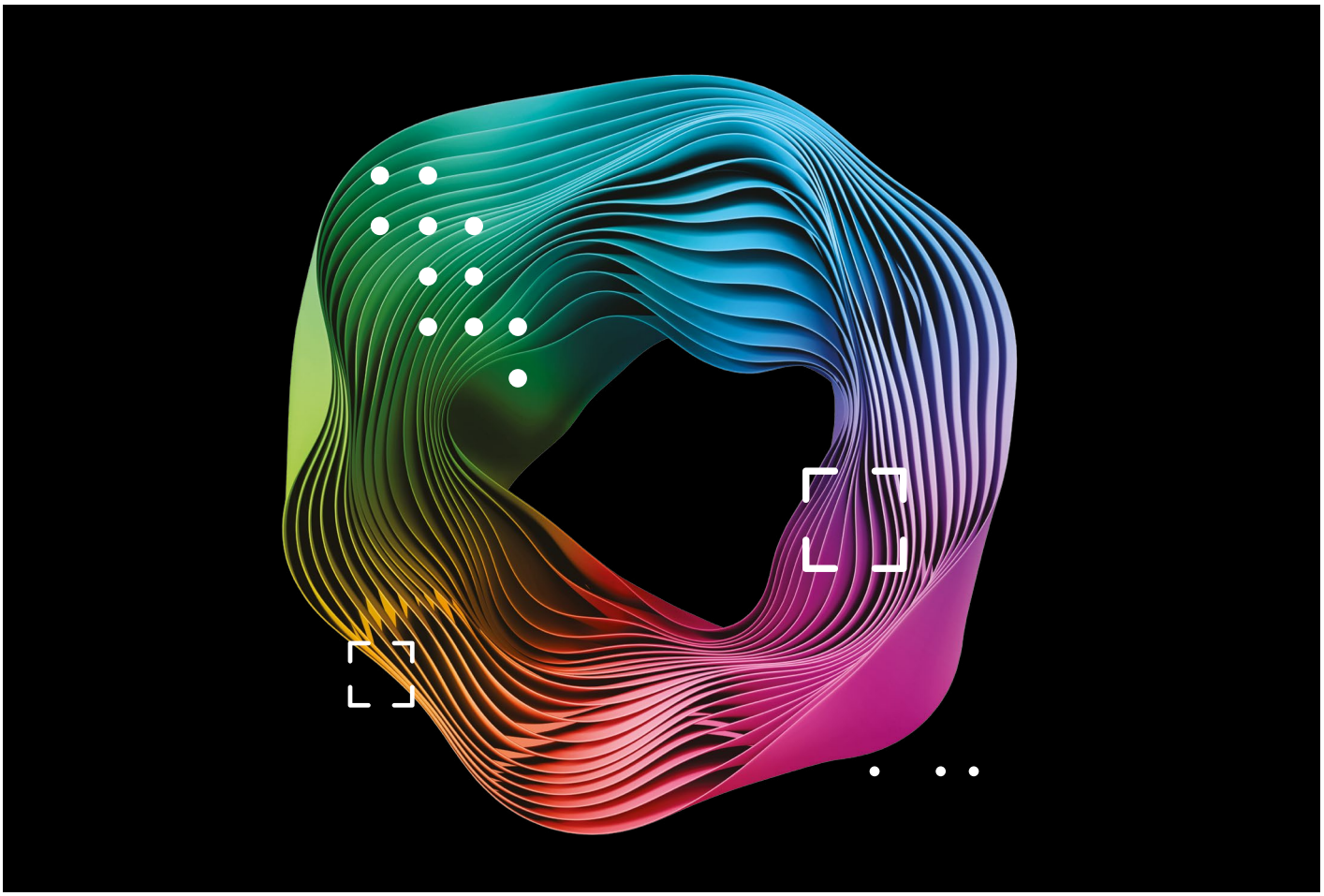
Deloitte.

Together makes progress



Agentic AI Governance – enhanced

Agentic AI Leadership Insights Series – Part 6

Agentic AI is integrating into enterprise workflows faster and more deeply than any previous wave of technology. The primary reason: While Generative AI merely recommends actions for humans to take, Agentic AI acts independently — driving unprecedented gains in speed, quality, and cost-efficiency. Yet it is precisely this agency that raises the stakes. A snarky chatbot is an inconvenience; an autonomous agent that triggers workflows, executes transactions, and orchestrates entire process chains introduces material operational risk. The question for senior leaders is how to unlock the efficiency of autonomous systems while keeping control of quality, security, ethics, and cost.

Inherent agentic risks

Many risks of agentic AI will sound familiar. AI is arguably more of a driver of risk than a distinct risk class. It puts an increasing amount of analytical and automation power into the hands of a wider group of people. This has given rise to the notion of "Shadow AI" risk: either employees using unsanctioned tools or applying approved tools in ways they were never intended. Another emerging threat is the poor design associated with "vibe coding" (building software by trial-and-error using AI without proper engineering rigor). In both cases, the risk stems from human misuse rather than the technology itself. In both cases, the risk lay less with the technology than its inappropriate use.

Agentic AI has, however, also introduced new risk surfaces, such as misalignment. Because reasoning engines interpret intent without requiring explicit instructions, they could misunderstand a goal, overgeneralize, or act on the wrong priority. Since most enterprises rely on third-party foundation models, they essentially operate these reasoning engines as "black boxes." This exposes them to supply-chain vulnerabilities, from sleeper exploits introduced during model tuning to prompt injection and model poisoning during live operations. A compromised model may execute differently than its user intends, acting on previous and hidden instructions. For instance, a malicious actor could exploit this vulnerability to exfiltrate proprietary data or suppress cybersecurity alerts. (For a deeper dive, see our previous article, "Five Fault Lines of Agentic AI.")

Complexity by design

What appears to users as a single "agent" is often a Multi-Agent System (MAS)—a dynamic team of specialized sub-agents, tools, and micro-services. While this architecture dramatically amplifies capabilities, its inherent variability introduces new failure modes. Complexity is introduced by the inherent interconnectedness of agentic systems with the systems and tool landscape they are charged with operating. Without clear status signals, checkpoints, and active monitoring, issues may go unnoticed until they have already caused material damage. For example, if a procurement agent misinterprets a budget threshold, a downstream scheduling agent might blindly approve dependent actions—propagating and compounding the error rather than containing it.

Proliferation risk

Managing multiple models across different systems is a significant lifecycle challenge. Low-code tools empower non-technical "citizen developers" to deploy agents rapidly. While a powerful productivity-booster, an unmanaged proliferation of agents can quickly create a control and security headache.[1] Agents with a given combination of system access rights could also unwittingly introduce security vulnerabilities. Furthermore, because reasoning engines are probabilistic, their behavior can drift over time or change abruptly when the underlying LLM is updated. A multi-agent system may rely on several third-party models. If model providers deprecate them, support ends and the agentic systems they power simply stop functioning. Inefficiently designed agents can also burn through tokens, leading to runaway operational costs (detailed in our next article, "The Token Machine: Runaway Agent Costs").

These and other factors make the governance of agentic AI an essential component of any agentic transformation.[2] Good governance ensures that Agentic AI is deployed responsibly, that issues are prevented where possible, that mitigation is rapid when needed, and that roles/privileges and accountability are clearly defined. It is the mechanism through which organizations uphold their obligations to clients, employees, and regulators. Effective agentic AI governance need not be bureaucratic. Instead, a lightweight, practical framework provides clarity, accelerates adoption, and drives ROI by successfully moving prototypes from proof-of-concept into full production.[3]

[1] Prakash et al., Agentic AI Governance and Lifecycle Management in Healthcare, 2026. This article addresses concerns of "agent sprawl" (duplication, unclear accountability, inconsistent controls, change in tool permissions) arguing the urgent need for rigorous controls in the form of agentic AI lifecycle management. https://arxiv.org/pdf/2601.15630

[2] Gartner, AI Ethics, Governance and Compliance, 2024. Gartner emphasizes that as AI systems gain autonomy, governance becomes critical to prevent loss of control, ensure accountability, and manage risks arising from misaligned or unconstrained AI behavior in enterprise environments. https://www.gartner.com/en/articles/ai-ethics-governance-and-compliance

[3] "AI Governance Maturity Assessment", Deloitte, Wurth et al, 2026

Agentic AI governance

"The best defense is a good offense." The principle of "action over reaction" applies as much in cyberspace, within your very own (fire)walls, as it does on the field. Under normal circumstances, agents are digital coworkers. When things go wrong, they can effectively behave as adversaries. In such situations, we concern ourselves with policies, KPI, early-warning indicators, control limits, oversight and remedial actions … in a word: governance.

The question is then: what constitutes good governance of agentic AI… a "good offense"? From the organizational to the technical, from how tools are built to how they are correctly used, agentic AI Governance covers a broad spectrum of activity. There are certainly common features to previous forms of governance — data governance, AI governance (machine learning, GenAI) — but agentic AI is sufficiently different in what it can do and how it works, that it merits closer inspection, starting with the three defining traits that set it apart:

- **Autonomy —** lightning fast and often silently in the background
- **Flexibility —** inferring intent rather than requiring explicit instruction
- **Complexity —** dynamically self-assembling components to solve a specific task

Unlike data governance, agentic AI governance must accommodate a highly dynamic, shifting playing field. Unlike classic AI governance focused on individual machine learning models, it is concerned with an unwieldy network of agentic components woven together differently for each task. This shifts the governance focus from static "use-case approval" to continuous, dynamic monitoring — a capability where legacy GRC platforms often fall short. Monitoring agentic systems requires a hybrid approach: blending software application monitoring with principles akin to human resource management. Yet even using leading indicators, monitoring agentic systems will always lag, they will always react, the damage already done. Proactive controls will become ever more important, building governance into the very design of the MAS.

While agentic governance will require new components at a detailed level, conceptually, they still fall neatly into the same three dimensions of AI governance (see our earlier publication "Operationalizing AI Governance"): oversight, quality, and risk management. They facilitate communication and conveniently align to the perspective of the EU AI Act. They do not suggest an organizational structure. The ideal agentic AI governance structure for individual organizations will differ, fitting into established structures and taking regional or sectoral nuances into account.

Generally, it is good practice to integrate AI risk into each function and maintain a lightweight coordinating level on top – ensuring both clear ownership and involvement of experts.

Put plainly, the three axes of agentic AI governance address the three central questions that confront the AI-aware leader:

- **Is a human clearly accountable for what the agents do?** → Oversight
- **Is the system built and tested well enough to be relied upon?** → Quality
- **Are we prepared for the cases where it still goes wrong?** → Risk Management

Agentic AI oversight

Oversight is about ultimate accountability – ensuring the structures are in place, the right people in the right jobs, and taking responsibility for eventual issues – and for fixing them. In that, oversight spans the entire lifecycle – from use case approvals through go/no-go deployment decisions – with an eye on upsides, risks, and total cost of ownership. It centers on human judgment and ultimate accountability. Many users have learned the hard way that AI outputs should not be blindly accepted. Some therefore conclude that agentic AI may never be relied upon to fully automate processes. That premature conclusion only robs agentic AI of its potential. A more effective approach is to introduce autonomy gradually, following rigorous lab and field testing, before and after deployment. A well-functioning agentic governance is as much process as it is infrastructure, as it follows a specific sequence (see table 1).

Agentic AI must earn trust rather than be granted it by default. Outputs must be reviewed by qualified staff. All business functions have a role to play. Senior leadership must treat AI as a strategic capability requiring ongoing attention, not a "set and forget" technology. Even after signing off full autonomy, AI leaders are never off the hook. The developers of an AI are responsible; their management is accountable. Always.

This is the central tenet of agentic AI oversight, and why it should appear on management agendas and be integrated into existing governance structures, alongside or as part of established frameworks for ethics, privacy, consumer protection, and product safety. Oversight begins with defining the KPI against which agent performance will be measured and ends with ownership for agentic outcomes. In practice, oversight means being explicit about what to look out for. Leaders need not master the technical implementation. However, they must insist that monitoring covers three critical failure modes:

- **Prompt Injection:** Agents manipulated by malicious instructions hidden in files or data.
- **Tool Misuse:** Unintended or unauthorized actions taken on connected enterprise systems.
- **Scope Creep:** Agents operating outside their approved operational boundaries.

Naming the specific, expected behavior up-front is what transforms oversight from a lofty principle into tangible control. Some might argue that this view of agentic AI governance is missing a fourth dimension: AI compliance. Our view is that compliance is a formalized consequence of sound governance, not its primary motivator. The same process-rigor and documentation required for quality and risk management also strengthens compliance across all levels of AI maturity.

Tab. 1 – A sequence for building trust

Design	Test	Monitor
Embed governance and safety guardrails into the system architecture from day one to prevent silent, rapid errors.	Conduct rigorous, simulation-based testing in sandboxed environments to expose hidden failure modes before deployment.	Run continuous, deep operational monitoring to catch and contain real-world drift or errors before they propagate.

Agentic AI quality

AI quality is about prevention, catching problems before they materialize. It is about ensuring a degree of governance by design … indeed also regulatory compliance by design. That includes standards, guardrails, rigorous testing … and stringent role-based access controls. For agentic systems, this means robust test cases, simulation environments, automated evaluations, A/B tests, and critic agents. Structured documentations — from model cards through to a roster of agentic systems in a single system of record — are indispensable to maintain quality control. After deployment, continuous monitoring, drift detection, and issue logging ensure that quality is not a one-time exercise. Regular monitoring, scheduled review cycles and even agent expiration dates are among the arsenal of tools that keep control over the "fleet" of agentic systems in operation. The most successful organizations define clear acceptance criteria for agentic workflows, so teams know exactly when a process is considered reliable enough for real-world deployment.

Designing governance into the MAS is a subset of building quality into the design of any AI, or indeed any IT application. The characteristics of good quality are conveniently summarized by the principles of the Deloitte Trustworthy AI Framework. Those principles and the metrics by which they are quantified apply to monitoring as much as they do to testing. (Monitoring is essentially continuous testing over time.)

- **Transparency/explainability.** To track performance, implement verbose logging of every action—forcing agents to declare their intent and log results. These "trace-routers" establish an audit trail that is invaluable for Human-in-the-Loop (HITL) oversight, and can be analyzed by a separate "LLM-as-a-judge" monitoring system. Telemetry lays the foundation for testing and monitoring.
- **Security/safety.** Role-Based Access Control (RBAC) should be strictly enforced, applying the principle of least privilege, including temporary, time-bound access rights that expire automatically upon task completion.
- **Data protection/privacy.** Reliance solely on pre-labeled data can be risky: it is better to integrate NLP driven pre-processing to scan and redact Personally Identifiable Information (PII) before it reaches the agent, and route highly sensitive processing to secure, on-premises environments.
- **Bias/fairness.** While bias can and should be measured during the iterative development of an AI solution, it is most prominently addressed during testing & monitoring. If bias is detected, it should be duly analyzed and corrected at its source—not "de-biased" (a bandage solution), but re-designed so that the bias does not arise in the first place.
- **Robustness/reliability.** In the development phase, robustness is topic of testing focused on accuracy of outputs vs. intentions, a gold-standard or ground truth. In operation, robustness becomes more a monitoring topic, focused on identifying deterioration in accuracy (data drift, concept shift) over time.
- **Regulatory and policy compliance.** Each agent's action or output should comply to company policies, as well as to relevant regulation such as the EU AI Act, GDPR, DORA … or local, sector-specific regulation (EU Machinery Regulation, CER, NIS2 …).

The concept of "quality by design" applies as much, if not more, to the "harness" (the software connection layer) than to the LLM "brain" of the agentic system itself. Tool use is one example at the very core of agentic AI. The master (orchestrator) agent ascertains which tools and sub-agents it needs to accomplish each task by searching a tool registry where their respective functionalities are described. This functions properly if attention is given to the accuracy, clarity and completeness of those tool/sub-agent functional descriptions. These and several further quality measures may be taken to avoid divergent agent behavior, for example:

- Precise (and narrow) descriptions of the agent's tasks and responsibilities
- Consistent instructions on how the agent should act
- Clear, concrete, explicit examples illustrating the agent's role in the system
- Mutually exclusive, non-overlapping roles/distinct boundaries, hand-offs (to avoid conflicts with other agents or governance components)
- Sufficient adaptability or responsiveness to changing circumstances
- Including feedback and learning mechanisms for continuous improvement and error correction

Comprehensive testing should encompass the integrity of the tools at the agent's disposal, in addition to examining the agents themselves. Tool use is arguably the single-most distinctive factor that makes agentic AI so compelling: identifying the right tools for the job, then operating them in a process or applying them to a dataset. Even if the agent's reasoning engine functions properly, the overall MAS may be vulnerable to a different failure mode if the tools the agent uses to get the job done are either faulty or broadcast to the agent incorrectly (overstated capabilities). Testing these tools in isolation—before integrating them into the multi-agent system—prevents unexpected behavioral failures.

Agentic AI risk management

While good quality will prevent many risks from materializing, no one can anticipate everything. AI risk management focuses on business continuity by preparing the organization for rapid and effective response to uncertainty. It anticipates potential failure modes, estimates likelihood, quantifies impact, and defines mitigation strategies. It also isolates residual risk so leaders can judge whether it is acceptable and make informed decisions. Decades of model risk management experience in regulated industries offer a strong foundation. In the development phase, a decision matrix mapping risk levels to mandatory oversight steps often helps teams evaluate the cost of risk management, an important factor in deciding which use cases are most suitable for agentic automation. Post deployment, risk management policies dictate the necessary review frequency of agents in operation, and how exceptions are to be escalated.

The biggest AI risk

The purpose of this article is to raise awareness for agentic AI risk, and to deliver the confident message that they can be managed well, with the right expertise and attention. Reading only this article in the series might leave the impression that we regard agentic AI negatively, something that must be tolerated and controlled. That could not be further from the truth. AI is the defining technology of our time and agentic AI is a significant milestone. We have only begun to tap its potential. Yes, there are risks associated with AI agents — at many levels. For the deployer, probabilistic AI systems may add new capabilities and flexibility, but they exit the comfort zone of deterministic systems to which we are accustomed — that give the same result (right or wrong) every time, in an easily explained way. Those systems were built on human rules. GenAI models — the brains of agentic AI — are built on vast amounts of data, containing knowledge and patterns at a scale no single human can digest. Yet the most significant risk is often left unaddressed: the risk of inaction. Avoiding or over-constraining agentic technology out of fear is a recipe for competitive obsolescence. Failing to harness the potential of agentic AI is not just a strategic misstep—for many industries, it represents an existential threat. Our message is clear: embrace agentic AI with eyes wide open. Do not shy away from the challenges; instead, capture the opportunity and master the risk with an efficient, low-overhead governance framework.

Priority moves to consider

- **Standardize Quality Gates:** Define clear test cases, performance baselines, and monitoring thresholds before any agent goes live.
- **Map Agentic Risks:** Maintain a dedicated risk register for agentic workflows with pre-defined mitigation strategies for high-impact failure modes.
- **Mandate Human Oversight:** Implement structured Human-in-the-Loop (HITL) checkpoints, prioritizing customer-facing and financial workflows.
- **Consolidate Frameworks:** Embed agentic governance into your existing enterprise risk management (ERM), privacy, and compliance structures to prevent bureaucratic duplication.
- **Assess Enterprise Readiness:** Conduct a rapid diagnostic of current tooling, technical skills, and leadership alignment to identify and close execution gaps.

Your contact

David Thogmartin
Partner
aiStudio, AI & Data Analytics,
Trustworthy AI
Tel: +49 151 58072163
dthogmartin@deloitte.de

Jan Hejtmánek
Partner
Intelligent Automation
Artificial Intelligence & Data
Tel: +420 731 685 523
jhejtmanek@deloittece.com

Deloitte.

Issue 07/2026

MAKING AN IMPACT THAT MATTERS since 1845