

Governing Agentic AI in Financial Institutions



This document introduces a governance-architectural contribution to the financial services literature on Generative and Agentic AI. It is intended as an overview of the accompanying whitepaper and handbook that explores the topic in greater depth. Its central innovation lies in demonstrating that the risks introduced by Agentic AI do not require the invention of a new, AI-specific governance paradigm. Instead, they require a more explicit, continuous, and disciplined application of existing [Model Risk Management \(MRM\) frameworks](#), fully integrated within the [Three Lines of Defence \(3LoD\)](#) and [explicitly connected to established Operational Risk Management \(ORM\) practices](#).

By systematically embedding agentic autonomy, behavioural complexity, tool use, and dynamic execution within established [tiering, lifecycle, validation, and accountability structures](#)—and by clearly allocating responsibilities across the first, second, and third lines of defence—the document provides institutions with a [supervisory-credible, cross-jurisdictional approach](#) to scaling Agentic AI while preserving control, auditability, and accountability. Crucially, it clarifies that while [Agentic AI risk is analytically governed through MRM](#), its failures [primarily materialise as operational risk events](#), affecting people, processes, systems, and external dependencies, and must therefore be [prevented, detected, escalated, and remediated through existing operational risk and incident management frameworks](#).

The framework and use cases demonstrate that Agentic AI can be deployed in financial institutions in a way that simultaneously

enhances business value, strengthens risk and operational resilience, and meets evolving supervisory expectations—[provided that autonomy is governed through an integrated MRM, ORM, and 3LoD approach](#), rather than through isolated technical or ethical overlays.

The key takeaways can be grouped into five interrelated dimensions:

1. Business Advantage: Scaling Intelligence Without Losing Control
2. Key Insights: Why Traditional AI Governance Is Insufficient for Agentic Systems
3. Innovation in Controls: Preventive, Detective, and Adaptive Governance at the Intersection of MRM and ORM within the 3LoD
4. Translation into Practice — Lessons from the Two Use Cases: an insurance scenario in which Risk acts as a user to explore complex scenarios and emerging risks within controlled, reviewable behavioural boundaries, and a banking scenario in which Risk acts as a steward to make autonomy, change, monitoring, and incidents auditable and enforceable
5. Regulatory Expectations, Best Practices, and Cross-Jurisdictional Gaps

1. Business Advantage: Scaling Intelligence Without Losing Control

The primary business advantage of the combined MRM and 3LoD framework is that it allows institutions to scale agentic capabilities [without sacrificing accountability, auditability, or decision ownership](#).

Agentic AI enables faster analysis, continuous decision support, and more adaptive responses across both banking and insurance processes. However, without a disciplined governance backbone, these same characteristics introduce unacceptable [operational, conduct, and prudential risks](#), as errors propagate through automated workflows, human oversight weakens, and system dependencies multiply. The framework demonstrates that business value does not come from autonomy alone, but from [controlled autonomy](#)—autonomy that is [explicitly bounded, benchmarked, observable, and accountable](#).

In the insurance use case, where the risk manager acts as a [user](#) of Agentic AI, the framework enables richer scenario analysis, forward-looking risk insights, and more efficient capital assessment processes, while preserving actuarial judgment and management accountability. In the banking use case, where the risk manager acts as a [steward](#), the framework supports the safe deployment of agentic systems within capital modelling, stress testing, and risk aggregation processes, ensuring that productivity gains do not undermine model integrity, operational resilience, or supervisory confidence.

In both cases, the business benefit arises precisely because Agentic AI is treated as a [risk-bearing system embedded in core governance and operational control frameworks](#), rather than as an experimental or peripheral technology.

2. Key Insights: Why Traditional AI Governance Is Insufficient for Agentic Systems

Agentic AI does not merely increase model complexity; it fundamentally changes how risk propagates through the organisation. Autonomy, multi-step reasoning, tool use, and delegation introduce dynamic and non-linear risk behaviours that cannot be governed through static approval processes or point-in-time validation alone.

The framework makes clear that traditional governance mechanisms—such as initial model approval, documentation, or periodic review—are necessary but not sufficient. Risk does not stabilise at model approval. Instead, it evolves continuously as agentic systems interact with data, tools, users, and external

dependencies. This explains why agentic systems may escalate into higher risk tiers over time even when their individual components appear benign in isolation, and why supervisors increasingly focus on controllability, explainability, behavioural discipline, and observable operational outcomes rather than on the novelty of the underlying technology.

The two enterprise use cases illustrate this dynamic clearly. In insurance, even agentic systems positioned as advisory tools can materially influence capital assessment, reserving decisions, and management actions, thereby requiring governance standards comparable to those applied to high-impact actuarial and risk models. In banking, agentic overlays embedded into existing risk engines or finance processes amplify usage intensity and downstream impact, triggering heightened supervisory expectations even where the formal regulatory classification of the underlying model has not changed.

[For these reasons, the document explicitly reframes AI governance as both a model risk and an operational risk problem, rather than a purely technological one](#). From a model risk perspective, agentic systems introduce structural deviation risk: the possibility that outputs, recommendations, or decisions diverge from appropriate benchmarks, challenger models, expert judgement, or expected economic behaviour. This risk must be identified, measured, and challenged through robust MRM practices already at the development, validation, and revalidation stages.

At the same time, agentic systems operate within live processes, systems, and organisational workflows, where risk materialises through execution rather than design alone. From an operational risk perspective, failures may arise from system malfunctions, control breakdowns, data issues, human-in-the-loop failures, misconfiguration, or unintended interactions across systems and third-party dependencies. These risks affect not only technical accuracy, but also operational resilience, customer outcomes, financial loss, and reputational integrity.

[By explicitly recognising both dimensions, the document moves AI governance away from a narrow focus on technology oversight and towards an integrated risk discipline. MRM defines acceptable behaviour and measurable deviation, while ORM ensures that agentic systems are deployed and operated within resilient, controlled environments](#).

This dual framing provides the conceptual foundation for the integrated control framework developed in the subsequent sections.

3. Innovation in Controls: Preventive, Detective, and Adaptive Governance at the Intersection of MRM and ORM within the 3LoD

The most substantive innovation of the framework does not lie in the introduction of new or AI-specific controls, but in the disciplined embedding of preventive, detective, and adaptive controls within existing MRM frameworks, explicitly anchored on a foundation of robust ORM controls and operationalised through the 3LoD.

Within this architecture, MRM and ORM play distinct but complementary roles. Governance effectiveness therefore depends on their combined application within the 3LoD, rather than on either discipline in isolation.

Preventive controls are primarily embedded through lifecycle governance and validation design. Controls such as prompt policies and approved prompt templates, safe action-space constraints, guardrail libraries, and explicitly designed human-in-the-loop checkpoints define the behavioural perimeter of agentic systems before deviation materialises. These controls are designed and operated by the first line, independently challenged by the second line, and assured by the third line. Their effectiveness relies on supporting operational controls, including disciplined process execution, access management, segregation of duties, and system and infrastructure resilience.

Detective controls operate through extended validation and continuous monitoring. Output validation, logging and observability, behavioural monitoring, and deviation analysis ensure that benchmark divergence, hallucinations, autonomy drift, or unauthorised delegation are detected early and assessed proportionately to model tier and use-case materiality. These signals are not treated as purely analytical artefacts; they feed directly into operational risk monitoring and incident management processes, enabling timely escalation, loss prevention, and corrective action.

Adaptive controls ensure that governance remains effective over time as agentic systems evolve. Independent change management, escalation mechanisms, and incident response processes allow institutions to respond to behavioural drift, foundation-model updates, configuration changes, and deployment extensions without erosion of control. In this context, ORM plays a central role in ensuring operational resilience, recognising that disruption is expected and that recovery capability is as critical as prevention.

Taken together, preventive, detective, and adaptive controls **do more than constrain technical model behaviour**. They **define the conditions under which agentic systems can be responsibly deployed in decision-critical environments**. This creates a natural bridge between traditional risk controls and internationally recognised responsible-AI principles focused on explainability, data integrity, governance discipline, and ethical outcomes.

A further innovation of the framework lies in how it incorporates internationally recognised responsible-AI principles—often articulated in terms of Explainability, Data, Governance, and Ethics (EDGE)—without introducing a parallel or standalone AI governance regime. Instead, these principles are translated into explicit, enforceable control expectations within the existing MRM and ORM architecture, fully embedded in the 3LoD.

Explainability and Governance are already central to the framework. Explainability is operationalised through tier-proportionate validation requirements, decision provenance, logging, and escalation thresholds, while governance is enforced through clear ownership, approval, and accountability structures across the lifecycle. **However, explainability and governance alone are insufficient once agentic systems operate autonomously and at scale. The more challenging dimensions—data risk and ethical risk—are therefore addressed as governable risk drivers that condition both model risk and operational risk outcomes.**

Data risk is treated as a foundational driver of agentic model risk rather than a purely technical concern. In agentic systems, deficiencies in data quality, lineage, representativeness, timeliness, or integrity are amplified by autonomous reasoning, retrieval-augmented generation, external tool use, and dynamic memory. Accordingly, data risk is governed through MRM controls assessing data suitability, benchmark relevance, and sensitivity at development and validation stage, and reinforced through ORM controls covering data management processes, access controls, third-party dependencies, and resilience against data disruption.

Ethical risk, including bias, fairness, and unintended discriminatory outcomes, is framed as an outcome-oriented risk dimension with prudential, conduct, and reputational implications. In agentic systems, ethical risk may emerge over time through adaptive behaviour, interaction effects, and changing operational contexts. Ethical considerations are therefore operationalised through measurable constraints,

testing protocols, and escalation thresholds embedded within MRM validation and ongoing monitoring, and enforced through ORM mechanisms such as accountability assignment, auditability, customer recourse, and incident escalation.

This approach reflects a growing regulatory consensus that responsible-AI principles must be embedded within existing risk governance frameworks rather than treated as parallel aspirational standards. The dialogue between regulators and industry that informed internationally recognised EDGE principles has also directly influenced supervisory expectations on MRM, including the explicit inclusion of artificial intelligence and machine learning within the scope of modern model risk guidance. [By translating EDGE principles into practical, auditable, and enforceable controls aligned with MRM, ORM, and the 3LoD, the framework ensures that responsible-AI objectives strengthen—rather than dilute—existing governance.](#)

At the same time, MRM and ORM remain conceptually distinct. In fact, meanwhile MRM evaluates structural deviation risk, ORM evaluates the likelihood and impact of loss-causing events arising from the execution of activities in which agentic systems operate. ORM therefore addresses how risks materialise operationally once agentic systems are embedded in real processes, systems, and organisational structures.

This positioning reflects established supervisory thinking that ORM is not limited to the retrospective analysis of loss events. [It includes explicit ex-ante preventive and resilience controls, disruption management, and recovery across critical operations.](#) Within the combined framework, MRM therefore rests on a foundation of strong operational risk controls, ensuring that model-level safeguards remain effective in real-world execution. As a result, forward-looking structural risk signals identified through MRM directly inform operational risk scenarios and control design. In this way, structural model risk is translated into practical, operationally effective prevention, mitigation, and recovery mechanisms.

Crucially, these controls are not standalone safeguards. They represent the operational expression of tiering outcomes, enforced through clearly defined responsibilities across the 3LoD and aligned simultaneously with model risk discipline and operational resilience objectives. In this way, institutions can address the specific risks introduced by Agentic AI without expanding the risk taxonomy or creating parallel governance structures, but by strengthening the coherence and execution of existing MRM and ORM frameworks.

4. Translation into Practice: Lessons from the Two Use Cases

The two enterprise use cases demonstrate how the same governance framework flexes across sectors while preserving a [common control logic](#).

In the insurance use case, the risk manager acting as a user benefits from agentic capabilities to explore complex scenarios, emerging risks, and capital sensitivities. The framework ensures that these insights are produced within [controlled behavioural boundaries](#), transparently documented, independently reviewable, and operationally anchored through existing incident and escalation processes.

In the banking use case, the risk manager acting as a steward focuses on ensuring that agentic systems embedded in critical risk and finance processes remain consistent with [risk appetite, regulatory expectations, and internal model governance](#). The framework enables this stewardship by making autonomy, change, monitoring, and operational incidents [auditable and enforceable](#).

Across both cases, the combined MRM, ORM, and 3LoD approach clarifies ownership despite increased autonomy, enables independent challenge without stifling innovation, and provides a defensible narrative for supervisors across jurisdictions.

5. Regulatory Expectations, Best Practices, and Cross-Jurisdictional Gaps

In the handbook it is provided a cross-jurisdictional regulatory analysis, consolidated in the Annex of that document. Here it is worth mentioning that Agentic AI is already materially regulated - albeit indirectly - through existing prudential, conduct, governance, and operational resilience frameworks and this is demonstrated by a structured comparison of regulatory obligations, industry best practices, and regulatory gaps across the EU, UK, US, and Canada. Through a structured comparison of regulatory obligations, industry best practices, and regulatory gaps across the EU, UK, US, and Canada, the document demonstrates that Agentic AI is [already materially regulated—albeit indirectly—through existing prudential, conduct, governance, and operational resilience frameworks](#).

The analysis shows that while no jurisdiction yet provides a complete, explicit rulebook for agentic autonomy, supervisory expectations consistently converge around governance outcomes: [accountability, controllability, explainability, proportionality, operational resilience, and auditability](#). Where regulation remains silent—particularly on multi-agent behaviour, dynamic delegation, and behavioural drift—institutions are expected to rely on [robust internal governance and operational risk frameworks](#), not permissive interpretation.

The combined MRM and 3LoD framework therefore functions not only as a governance model, but as a [regulatory and operational resilience strategy](#), enabling institutions to bridge regulatory gaps using documented best practices and to engage credibly with supervisors across jurisdictions.

Concluding Takeaway

The overarching takeaway is that [Agentic AI does not require a new governance paradigm](#). It requires that existing governance—Model Risk Management, Operational Risk Management, and the Three Lines of Defence—be applied more rigorously, [more continuously, and more explicitly](#), with clear distinctions between [ex-ante deviation control](#) and [ex-post loss and resilience management](#).

This document is complemented by a white paper "**Governing Agentic AI in Financial Institutions - Integrating Model Risk Management, Operational Risk Management and the Three Lines of Defence into a Holistic Control Framework**", which provides a deeper conceptual and analytical exploration of the underlying governance principles, and by a handbook "**Governing Agentic AI in Financial Institutions - Integrating Model Risk Management, Operational Risk and the Three Lines of Defence into a Holistic Control Framework for Regulatory and Best Practice Compliance: Practical Use Cases for Risk Managers**", which offers a more detailed and practically oriented treatment of the same topics. Together, they show how the combined framework translates from principle to practice in a way that is [scalable, auditable, operationally resilient, and supervisory-credible](#), while remaining fully aligned with established regulatory thinking across banking and insurance.

About the authors



Mario Onorato
Promontory Director
IBM Industry Diamond
monorato@promontory.com



Orazio Lascala
Promontory Director
olascala@promontory.com



Silvia Peschiera
Strategy and Transformation
Service Line Leader, Italy
speschiera@it.ibm.com



Silvia Procacci
Enterprise Strategy Practice
Leader, Italy
silvia.procacci@ibm.com



Giulia Ugolini
Managing Consultant, Italy
giulia.ugolini@it.ibm.com



Riccardo Cinelli
Managing Consultant, Italy
riccardo.a.cinelli@it.ibm.com



Matteo Muscolo
Strategy Consultant
matteo.muscolo@ibm.com

© Copyright IBM Corporation 2026

Produced in the
United States of America
February 2026

IBM, the IBM logo, and IBM Trademarks List, are trademarks or registered trademarks of International Business Machines Corporation, in the United States and/or other countries. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/trademark.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

