

Governing Agentic AI in Financial Institutions

Integrating Model Risk Management, Operational Risk
Management and the Three Lines of Defence into a
Holistic Control Framework

A White Paper



Foreword

As AI technologies advance, financial institutions are increasingly adopting agentic systems that can act autonomously, analyse complex data, and generate insights with minimal human intervention. While the potential for efficiency, innovation, and enhanced decision-making is enormous, these capabilities also introduce new dimensions of operational, model, and regulatory risk.

This report is a timely and invaluable contribution to the field. By bridging Model Risk Management, the Three Lines of Defence, and regulatory expectations, it presents a holistic and practical framework for managing the full lifecycle of agentic AI. The call for a mature Operational Risk Management foundation within the Three Lines of Defence framework to complement and in fact enable effective Model Risk Management resonates strongly from a practitioner's perspective. Particularly noteworthy are the use cases that illustrate the dual role of risk managers—as both active users and vigilant stewards—reinforcing the importance of explainability, data stewardship, governance and oversight, and ethical application in practice.

I am confident that this report will become a key reference for risk professionals, auditors, and senior management alike. It provides not only the theoretical foundations but also the practical tools and examples necessary to navigate the complex regulatory landscape, implement robust controls, and foster responsible innovation in AI.

Dr. Michael Zerbs,

Group Head (retired), Technology and Operations at Scotiabank
Director, Global Risk Institute and Unity Health

Strategic Perspectives

In recent years, the rapid evolution of Artificial Intelligence has presented both unprecedented opportunities and complex challenges for financial institutions. Agentic AI—capable of autonomous decision-making and multi-step reasoning—requires a careful balance between innovation, control, and accountability.

This report was conceived to provide a practical, holistic framework for governing AI in financial institutions. Unlike traditional approaches that treat Model Risk Management, the Three Lines of Defence, and regulatory compliance in isolation, this report integrates these perspectives to offer actionable guidance for risk managers. The inclusion of real-world use cases demonstrates how risk professionals can act both as users and stewards of AI models, ensuring alignment with organisational objectives, risk appetite, and supervisory expectations.

It is my hope that this report will serve as both a reference and a roadmap—helping financial institutions adopt AI responsibly, strengthen oversight, and embed risk-aware practices throughout the model lifecycle. By combining regulatory expectations, industry best practices, and hands-on examples, we aim to equip risk managers with the tools and insight needed to navigate this evolving landscape confidently.

Shanker Ramamurthy

Global Managing Partner, Banking & Financial Markets

IBM Consulting

Table of contents

	1. Introduction	17	6.3 Key Innovative Element of the “Combined” Framework
7	2. Industry Context and Scope of Application	18	6.4 Integration of Agentic AI Governance with Operational Risk Frameworks
	2.1 Industry Context		
7	2.2 Scope of Application	20	6.5 MRM and ORM as Complementary Control Frameworks for Agentic AI
	3. Cross-Jurisdictional Comparison and Emerging Supervisory Gaps	22	6.6 Key Takeaways of the Holistic MRM, Operational Risk and 3LoD Framework
9	3.1 Overarching Legal and Supervisory Landscape		
9	3.2 Banking Sector – Global Perspective	23	7. Enterprise Use Cases: From Governance Principles to Operational Value
10	3.3 Insurance Sector – Global Perspective		7.1 Insurance Use Case: Risk Management as User of Agentic AI
10	3.4 Cross-Sector Synthesis	24	7.2 Banking Use Case: Risk Management as Steward of Agentic AI
	4. The Dual Role of Risk Management in a GenAI & Agentic AI World	25	7.3 Synthesis: A Unified Framework Across Roles and Sectors
11	4.1 Risk Management as User of Agentic AI		
12	4.2 Renewal of Policies, Standards, and MRM Frameworks	26	8. Key Insights
12	4.3 Risk Management as Steward of Enterprise Agentic AI		8.1 Business Advantage: Scaling Intelligence Without Losing Control
	5. Methodological Implications: Embedding Agentic AI into MRM	27	8.2 Key Insights: Why Traditional AI Governance Is Insufficient for Agentic Systems
13	5.1 Revised Model Inventory		
13	5.2 A Tier-Based Model Risk Perspective	27	8.3 Innovation in Controls: Preventive, Detective, and Adaptive Governance at the Intersection of MRM and Operational Risk within the 3LoD
13	5.3 Methodological implications		
	6. Integrating Model Risk Management, Operational Risk Management and the Three Line of Defence	29	8.4 Translation into Practice: Lessons from the Two Use Cases
15	6.1 Implications for the Banking Industry		
16	6.2 Implications for the Insurance Industry	29	8.5 Regulatory Expectations, Best Practices, and Cross-Jurisdictional Gaps
			9. References

Introduction

This white paper makes a governance-architectural contribution to the financial services literature on Generative and Agentic Artificial Intelligence (AI).

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 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 white paper 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, Operational Risk, and Three Lines of Defence approach, rather than through isolated technical or ethical overlays.

This white paper is organized into six main sections that guide the reader from context and regulatory expectations to practical implementation and supervisory alignment.

- **Section 2 – Industry Context:** Examines the banking and insurance sectors, defining the scope of material prudential, model and operational risks, and highlighting sector-specific considerations and key takeaways.
- **Section 3 – Cross-Jurisdictional Comparison:** Presents a comparative analysis of emerging supervisory expectations and regulatory gaps across geographies and sectors, providing a foundation for harmonized governance approaches.
- **Section 4 – The Dual Role of Risk Management:** Explores risk management both as a user and as a steward of agentic AI systems, addressing methodological implications, regulatory differences, and unified but context-aware responsibilities.

- **Section 5 – Model Risk Management for Agentic Systems:** Details tiered model inventory, risk categorization, validation, and control design, establishing the structural logic for governing agentic behavior and emergent risks.
- **Section 6 – Integrating MRM and the Three Lines of Defence:** Expands the framework to incorporate operational risk domains, mapping model-specific risks to people, process, system, and external-event risks. This section demonstrates how tiering, preventive and adaptive controls, and continuous monitoring create a scalable, auditable, and supervisory-credible governance system.
- **Section 7 – Enterprise Use Cases and Deployment Blueprint:** Provides practical illustrations of risk management roles as both users and stewards of agentic AI, showcasing application of the combined framework across workflows, validation processes, human-in-the-loop interventions, and scenario testing.
- **Section 8 – Key Takeaways:** span five interconnected dimensions, from the business value of scaling agentic intelligence under control, to the limits of traditional AI governance, the evolution of preventive, detective and adaptive controls within the Three Lines of Defence, their practical application through real use cases, and the resulting regulatory expectations, best practices and cross-jurisdictional gaps.

Together, these sections show how the combined framework translates principles into practice, delivering governance that is **operationally resilient, auditable, and aligned with supervisory expectations**, while remaining adaptable across sectors, use cases, and organizational contexts.

Notably, while the formal scope of application in this paper is limited to prudential and compliance-relevant use cases, the underlying governance constructs — including the Three Lines of Defence, model and behavioral validation, lifecycle oversight, and incident management — reflect industry best practices that are equally applicable to the finance function and other materially decision-critical areas.

Industry Context and Scope of Application

2.1 Industry Context

There is a critical shift in the risk landscape driven by Generative AI (GenAI) and, more decisively, Agentic AI systems. Unlike traditional analytical or predictive models, agentic systems do not merely produce outputs; they plan, decide, act, and adapt within real operational environments. This transition marks a structural inflection point for financial institutions, as risk no longer resides solely in model accuracy or data quality, but in autonomy, interaction, and execution.

From an industry perspective, this evolution challenges long-standing assumptions embedded in governance, validation, and control frameworks. Agentic AI systems blur the boundaries between models, processes, and operational decision-making, making them materially different from both classical models and earlier generations of AI. The white paper positions this shift not as a future concern, but as a current supervisory and governance issue, with early deployments already influencing credit decisions, underwriting, claims handling, trading support, compliance analysis, and risk management workflows.

Crucially, the industry context is global. Financial institutions across the EU, UK, US, and Canada are experimenting with or deploying agentic systems, often using shared foundation models and cloud-based infrastructures. This creates systemic interdependencies and raises questions about concentration risk, third-party reliance, and cross-jurisdictional consistency.

2.2 Scope of Application

The scope of this paper is deliberately broad and cross-sectoral. It applies to:

- Banking institutions (retail, wholesale, investment banking).
- Insurance undertakings (life, non-life, health).
- Financial conglomerates operating across sectors and jurisdictions.

The analysis explicitly spans:

- EU regulatory regimes (AI Act, Capital Requirements Regulation (CRR) / Capital Requirements Directive (CRD), Solvency II).
- UK supervisory expectations (Prudential Regulation Authority (PRA), Financial Conduct Authority (FCA)).
- US Supervisory Guidance on Model Risk Management (SR 11-7–aligned practices).
- Canadian prudential frameworks (Office of the Superintendent of Financial Institutions (OSFI)).

Rather than proposing an AI-specific governance regime, the white paper frames Agentic AI within existing MRM disciplines, extending them where necessary to account for new behavioural and operational risks.

Banking benefits from explicit and mature MRM frameworks, supervisory manuals, and validation expectations that can be directly extended to agentic AI.

Insurance operates under a principles-based regime, where model risk is embedded across internal governance policies and procedures, actuarial standards, and general prudential requirements, but no standalone MRM regulation exists. As a result, insurers must rely more heavily on clear ownership, continuous monitoring, and demonstrable control of outcomes, rather than formal compliance with prescriptive validation or tiering rules.

[This makes the effective functioning of all three lines of defense – especially the first and second lines—structurally more critical in insurance than in banking when governing agentic AI.](#)

For clarity and consistency, this paper develops its methodology and governance framework within the defined scope of prudential risk and regulatory compliance. The focus is therefore on AI and model-enabled systems that are used in activities with direct implications for capital adequacy, solvency, liquidity, financial stability, supervisory reporting, and regulated decision-making. At the same time, the framework is intentionally designed to be principle-based and extensible.

In particular, finance activities such as valuation, reserving, financial reporting, planning, and performance management often exhibit risk characteristics comparable to prudential use cases, including high reliance on models, material judgment, and external reliance on outputs. As such, institutions may appropriately extend the framework described in this paper beyond the prudential perimeter, adapting its controls and proportionality to the specific risk profile and regulatory context of finance and business decision-support applications.

Cross-Jurisdictional Comparison and Emerging Supervisory Gaps

3.1 Overarching Legal and Supervisory Landscape

The second section undertakes a comparative analysis of how agentic AI is implicitly or explicitly addressed across jurisdictions. A key insight is that [no major jurisdiction yet regulates “Agentic AI” as a standalone category](#), but all materially regulate its effects through existing legal and prudential frameworks.

In the EU, the AI Act introduces a horizontal, risk-based classification of AI systems, explicitly capturing high-risk use cases common in financial services. While the AI Act does not use the term “agentic”, its requirements on human oversight, monitoring, and controllability map directly onto agentic risks. Importantly, these requirements coexist with sectoral regimes such as CRR/CRD and Solvency II, creating overlapping but reinforcing obligations.

In contrast, the UK, US, and Canada adopt [sector-specific, principles-based approaches](#), relying on supervisory guidance, operational resilience expectations, and MRM standards rather than horizontal AI legislation.

3.2 Banking Sector – Global Perspective

In banking, agentic AI systems frequently intersect with [high-impact model purposes](#), including credit decisioning, capital planning, stress testing, pricing, and liquidity management. Across jurisdictions:

- EU banks face explicit AI Act obligations layered on top of internal model governance.
- UK banks are governed by PRA/FCA expectations on model risk, accountability, and operational resilience.
- US banks rely on SR 11-7 and interagency MRM guidance, which partially—but not fully—addresses agentic behaviour.
- Canadian banks operate under OSFI Guideline E-23, which emphasises governance and resilience but lacks explicit agent-level controls.

A recurring supervisory gap is runtime behaviour: most frameworks assume models are static between reviews, while agentic systems adapt continuously.

3.3 Insurance Sector – Global Perspective

In insurance, agentic AI is emerging in actuarial modelling, underwriting, claims management, and fraud detection. While insurance regimes differ in capital mechanics, governance expectations around model risk, policyholder protection, and supervisory transparency are structurally similar.

The following considerations emerge:

- Solvency II provides strong internal model governance but was not designed for autonomous agents.
- UK and North American regimes rely on principles-based expectations rather than formal tiering.
- Supervisors increasingly expect insurers to demonstrate [explainability, auditability, and control](#), regardless of whether models are formally “regulatory.”

3.4 Cross-Sector Synthesis

Across all jurisdictions and sectors, this paper identifies a consistent pattern:

- Supervisors focus on [governance outcomes](#), not AI taxonomy.
- Accountability, controllability, and explainability are central.
- Existing frameworks can address agentic risks—but only if [applied more rigorously and continuously](#).

Summarising, the EU Artificial Intelligence Act establishes a binding horizontal regulatory baseline for high-risk AI systems across the financial sector, introducing requirements for classification, documentation, risk management, human oversight, post-deployment monitoring, and incident reporting. These obligations complement sectoral prudential frameworks such as Solvency II and CRR/CRD, forming minimum supervisory expectations for AI deployment.

Outside the EU, AI oversight relies on domestic supervisory frameworks, including UK prudential and conduct guidance, US model risk and supervisory standards, and Canadian governance expectations. While broadly aligned with EU principles, these frameworks remain jurisdiction-specific and reflect different legal and supervisory priorities. Agentic AI creates shared global challenges, including autonomous decision-making, multi-agent orchestration, third-party dependencies, emergent behaviours, and continuous monitoring requirements. A key supervisory priority across jurisdictions is translating ongoing monitoring into concrete governance and regulatory actions. Because existing financial regulations only partially address these risks, a comparative approach helps institutions distinguish between binding obligations, industry practices, and areas of regulatory uncertainty, supporting stronger AI governance and risk management across jurisdictions. All these information are provided in the handbook¹⁹.

The Dual Role of Risk Management in a GenAI & Agentic AI World

The adoption of Generative AI and Agentic AI at scale forces financial institutions, including both banking groups and insurance undertakings, to rethink the foundations of their enterprise control architecture. Risk Management is confronted with a dual imperative: it must govern the risks introduced by increasingly autonomous digital systems while simultaneously using those same systems to perform its own processes more efficiently.

4.1 Risk Management as User of Agentic AI

In both banking and insurance, second-line activities still rely heavily on manual synthesis of supervisory guidance, regulatory texts, internal policies, and complex risk data. These tasks are pervasive across geographies, whether institutions operate under the ECB/EBA in Europe, the PRA/FCA in the UK, the OCC/FRB/FDIC/CFPB in the US, or OSFI in Canada.

Agentic AI offers a similar value proposition across all these jurisdictions: it automates the most labour-intensive parts of risk analytics, documentation, and consistency checks.

Risk Management functions are no longer purely control bodies; they are also [users and beneficiaries](#) of Agentic AI. Use cases include:

- Scenario generation.
- Risk aggregation.
- Regulatory interpretation.
- Early warning signal detection.

This duality introduces a structural tension: Risk Management must rely on agentic systems while simultaneously [challenging and governing them](#).

4.2 Renewal of Policies, Standards, and MRM Frameworks

To manage this tension, this paper argues for a renewal—not replacement—of policies and standards. The adoption of agentic systems requires financial institutions to update existing policy and governance frameworks, as traditional model risk approaches are insufficient to address new AI behaviours and failure modes. Model risk policies, AI governance frameworks, and internal governance standards must be revised to align with evolving supervisory expectations, which consistently emphasise robustness, transparency, traceability, and human oversight.

Institutions must introduce new validation approaches covering prompts, agent workflows, adaptive learning mechanisms, and distributed decision-making, while expanding risk taxonomies to address autonomy risk, agent misalignment, synthetic data risk, multi-agent coordination failures, and concentration on foundation model providers. As a result, existing MRM frameworks remain valid but must be:

- Expanded to cover behavioural risk.
- Applied dynamically rather than episodically.
- Integrated with operational risk and ICT governance.

4.3 Risk Management as Steward of Enterprise Agentic AI

As banks and insurers begin integrating agentic AI into underwriting, claims, pricing, credit decisioning, transaction monitoring, fraud detection, customer engagement, and operational processes, the stewardship role of Risk becomes central. This means extending the MRM framework and updating governance standards to cover AI with varying degrees of autonomy, decision latitude, and interaction with enterprise systems.

As steward, Risk Management must:

- Define tiering methodologies that capture autonomy and impact.
- Set validation standards for behavioural stability.
- Ensure consistent application across business lines and entities.

This stewardship role is particularly critical in large, federated organisations where agentic systems may diverge across deployments.

In this dual role, Risk becomes both a [power user](#) of agentic AI and the architect of enterprise AI governance. Despite varying regulatory styles across the EU, UK, US, and Canada—and the different supervisory intensity between banking and insurance—the underlying responsibilities converge.

Risk must ensure that agentic AI enhances productivity, decision quality, and oversight capability, [without compromising resilience, regulatory compliance, or long-term strategic integrity](#).

Methodological Implications: Embedding Agentic AI into MRM

MRM is the primary governance anchor for Agentic AI systems, extending established model risk principles to account for autonomy, adaptive behaviour, and real-world execution. This section demonstrates that Agentic AI can be governed using the same disciplined, risk-based logic applied to other high-impact models, without creating parallel AI-specific governance regimes.

5.1 Revised Model Inventory

The introduction of a revised model inventory framework capable of capturing Agentic AI systems is composite, evolving model constructs rather than static analytical tools. Agentic systems are explicitly recognised as combinations of models, agents, tools, and workflows whose behaviour emerges over time. This revised inventory ensures traceability, ownership, and accountability across the full agentic architecture.

5.2 A Tier-Based Model Risk Perspective

Tiering is the foundational mechanism through which Agentic AI risk is translated into proportionate governance expectations. Tiering is driven by materiality, autonomy, decision impact, and external exposure, rather than by technical complexity alone.

Agentic-specific metadata ensures that autonomy level, tool-use authority, learning capability, and human override mechanisms are explicitly captured.

The second line performs structured challenge to ensure consistency, comparability, and risk-based calibration across agentic use cases.

Formal confirmation reinforces first-line accountability for intended use, limitations, and residual risks.

Tier assignment acts as the governance hinge, directly determining validation depth, monitoring intensity, escalation thresholds, and human-in-the-loop requirements.

5.3 Methodological Implications

Tiering outcomes are operationalised through preventive, detective, and adaptive controls embedded across validation, lifecycle management, monitoring, and change processes. Agentic AI risk is therefore governed continuously, not only at approval.

The objective is to assess agentic systems using established risk-based methodologies while explicitly recognising the new risk transmission mechanisms introduced by autonomy, emergent behaviours, delegated decision-making, and multi-agent interactions.

The methodological foundation remains the traditional tiering framework, which evaluates models based on materiality, model usage, and external impact. However, Agentic AI amplifies each

of these dimensions. Materiality increases as agent outputs increasingly influence financial performance, capital, pricing, and client outcomes through distributed and multi-step decision chains. This is particularly relevant where agents decompose objectives into sub-tasks and exchange intermediate outputs, creating heightened exposure to hallucination and error propagation across interconnected workflows.

Model usage is expanded when agentic systems move from advisory tools to operational decision-makers embedded within business processes. These systems often operate continuously, interact with multiple internal and external data sources, and dynamically delegate tasks across agents. As agent behaviour can evolve during runtime, usage risk becomes dynamic rather than static, increasing reliance on ongoing monitoring and governance escalation.

External impact grows when agent outputs affect customers or public disclosures, especially in cross-entity or federated deployments. Divergent model behaviour across entities may undermine supervisory expectations for consistency and comparability. Additional vulnerabilities arise from reliance on external data pipelines, shared training datasets, and third-party foundation model providers, which introduce exposure to data manipulation, vendor dependence, and systemic concentration risk.

These amplification mechanisms explain why Agentic AI models frequently escalate into higher risk tiers, even when individual analytical components appear low-risk. Tiering remains risk-driven rather than label-driven, and higher-tier classification automatically triggers stronger validation, governance, documentation, and management oversight requirements.

To ensure transparency and consistency in tiering decisions, the methodology identifies specific risk domains through which agentic behaviours amplify model risk. These include hallucination and fabrication risks, autonomy drift, delegation escalation, reinforcement learning instabilities, cross-entity model divergence, data poisoning, third-party dependence, concentration risk in foundation models, controllability limitations, and regulatory non-compliance risks. Collectively, these risks reflect the distributed and evolving nature of agentic systems, where accountability, causality, and control are inherently more complex than in traditional models.

While tiering determines the criticality of agentic systems, it does not by itself ensure effective risk control. The subsequent operational framework translates tiering outcomes into lifecycle governance, validation requirements, and supervisory oversight mechanisms. Within the 3LoD model, responsibilities are allocated across business, risk, and internal audit functions to ensure that Agentic AI governance is operationally enforceable, auditable, and aligned with industry and regulatory expectations.

Integrating MRM, Operational Risk and the Three Line of Defence

MRM framework operates in practice once Agentic AI systems are deployed within real organisations. While MRM provides the analytical and forward-looking discipline required to govern autonomy, behavioural deviation, and emergence, it is the integration with ORM and the 3LoD that makes governance effective, resilient, and supervisory-credible.

This section demonstrates that Agentic AI governance does not sit in isolation within model governance functions. Instead, it is embedded into existing accountability structures, operational processes, and assurance mechanisms across banking and insurance institutions. This integration ensures that agentic risks are both prevented *ex ante* and managed effectively when they materialise in operations.

6.1 Implications for the Banking Industry

For banks, Agentic AI governance must align with prudential expectations on model risk, operational resilience, and internal control frameworks. This section illustrates how the combined MRM–3LoD framework integrates seamlessly with established banking governance arrangements, avoiding the creation of AI-specific silos while addressing the heightened risks introduced by autonomy and execution capability.

The framework ensures that agentic use cases affecting credit decisions, financial reporting, regulatory submissions, or customer outcomes are governed with the same rigour as other high-impact models, while explicitly recognising their additional operational risk footprint.

Tiering Considerations and the First Line of Defence

Within the banking context, tiering directly shapes first-line responsibilities. The first line owns Agentic AI systems end-to-end, including their operational deployment, decision use, and outcomes. Tier assignment determines the intensity of controls embedded into business processes, such as human-in-the-loop checkpoints, approval thresholds, and runtime constraints.

This approach ensures that agentic behaviour is controlled at the point of execution, where risks materialise, rather than relying solely on *ex-post* review. The first line therefore remains fully accountable for both model-driven decisions and their operational consequences.

Validation and Revalidation Implications and the Second Line of Defence

The second line plays a critical role in independently challenging agentic systems throughout their lifecycle. In addition to traditional validation activities, the second line assesses

behavioural stability, benchmarking against market or human outcomes, and the adequacy of controls governing autonomy and delegation.

Revalidation is no longer a periodic compliance exercise but a dynamic process triggered by behavioural drift, scope changes, or evolving use patterns. This ensures that emerging risks are identified early and that first-line controls remain proportionate to actual risk exposure.

Governance and Accountability and the Third Line of Defence

The third line provides independent assurance that the integrated framework operates as intended. This includes assessing whether tiering decisions are consistently applied, whether validation and monitoring controls are effective, and whether accountability remains clear across organisational boundaries.



In the context of Agentic AI, third-line assurance focuses not only on control design but also on whether governance mechanisms genuinely prevent risk crystallisation rather than merely documenting it.

6.2 Implications for the Insurance Industry

The same integrated governance logic applies to insurance institutions, while aligning with International Association of Insurance Supervisors (IAIS) standards and Solvency II governance principles. Agentic AI use cases affecting underwriting, claims management, reserving, or customer interaction must be governed in a manner that ensures prudent decision-making and fair treatment of policyholders.

The framework ensures that agentic systems are treated as material decision-support or decision-making models, subject to proportional governance based on impact and autonomy.

Tiering Considerations and the First Line of Defence

In insurance, tiering determines how agentic systems are embedded into underwriting and claims processes. First-line ownership ensures that automated decisions remain within approved boundaries and that human judgment is retained where material uncertainty or customer impact exists.

Tier-driven controls prevent uncontrolled reliance on automation and ensure that accountability for outcomes remains clearly assigned.

Validation and Revalidation Implications and the Second Line of Defence

The second line independently challenges assumptions, behavioural outcomes, and fairness considerations. Validation explicitly considers whether agentic decisions align with actuarial benchmarks, historical outcomes, or expert judgment.

Revalidation addresses adaptive behaviour and data drift, ensuring that agentic systems remain fit for purpose as conditions change.

Governance and Accountability and the Third Line of Defence

Third-line assurance confirms that governance structures protect policyholders and support supervisory expectations. This includes verifying that escalation mechanisms, documentation, and human oversight operate effectively in practice, not only on paper.

6.3 Key Innovative Element of the “Combined” Framework

The transformation of MRM from a predominantly model-centric approval discipline into a [tier-driven, lifecycle-wide control system](#), operationalised through the 3LoD is the core innovation related to the governance of Agentic AI.

The combined framework does not introduce new AI-specific governance layers. Instead, it [re-engineers existing MRM and 3LoD mechanisms](#) so that they remain effective in the presence of agentic characteristics such as autonomy, behavioural adaptation, tool use, and real-world execution. The innovation lies in how tiering, validation, monitoring, and assurance are [reconnected into a single, coherent control logic](#) that operates continuously rather than episodically.

In this framework, tiering is no longer a static classification exercise. It becomes the [primary control lever](#) that determines how much assurance is required, how the lifecycle is governed, how validation is extended, and how monitoring and incident response are structured. The 3LoD provide the organisational mechanism through which this tier-driven logic is enforced in practice.

Tier-Driven Assurance: Validation, Documentation and Supervisory Review

Traditional assurance models assume relatively stable model behaviour between periodic reviews. Agentic systems invalidate this assumption by design, as they may adapt, delegate, or interact dynamically with external tools and data.

Tier-driven assurance ensures that the [depth, frequency, and formality of validation and documentation scale with risk](#). Higher-tier Agentic AI systems are subject to enhanced validation, including behavioural testing, explainability requirements, and explicit documentation of autonomy boundaries and escalation triggers. Lower-tier systems remain subject to proportionate controls, preserving operational efficiency.

From a supervisory perspective, this approach ensures that documentation and artefacts are not produced for compliance alone, but are [decision-useful](#). Validation reports, explainability artefacts, and lifecycle records are designed to support supervisory review, internal challenge, and audit, even where system behaviour evolves over time.

Model Lifecycle Redesign: Preventive Control of Agentic Behavior Over Time

In the combined framework, lifecycle management is no longer centred on post-deployment monitoring alone, but on [preventive control of behaviour over time](#).

For Agentic AI, risk often materialises not because the initial design was flawed, but because the system’s behaviour evolves in unintended ways once deployed. The lifecycle is therefore redesigned to embed controls ex ante, including constraints on autonomy, delegation limits, change approval mechanisms, and re-certification triggers linked to observed behaviour.

This redesign ensures that lifecycle governance becomes an [active risk management process](#), capable of intervening before deviations crystallise into losses or supervisory findings. Preventive lifecycle controls are a defining feature of the combined framework and a key differentiator from traditional model governance approaches.

Validation Extension and Independent Challenge of Agentic Risk

Traditional validation focuses on conceptual soundness, input-output relationships, and statistical performance. While these remain necessary, they are insufficient for Agentic AI.

Validation is therefore extended to assess [behavioural stability, interaction effects, and deviation from benchmarks over time](#), including scenarios involving tool use, delegation, and adaptive learning. Crucially, MRM remains anchored in [benchmarking and deviation analysis](#): risk is defined as the potential for material divergence from expected or acceptable outcomes, not merely as technical uncertainty.

Independent challenge by the second line plays a critical role in preventing normalisation of complexity. The framework explicitly guards against the risk that emergent behaviour or opacity is accepted as unavoidable. Instead, complexity increases the burden of justification, documentation, and constraint.

Continuous Monitoring, Incident Management and Adaptive Control

Within Agentic AI environments, risk crystallisation typically occurs in an operational and immediate manner rather than through gradual statistical deterioration. Because agentic systems dynamically interact with external tools, data sources,

and decision workflows, continuous monitoring and incident management represent the final governance layer through which tier-based controls are enforced once systems are deployed.

Effective Agentic AI governance relies on a coordinated system of preventive, detective, and adaptive controls operating across the full lifecycle. These controls should be viewed as an integrated, tier-driven framework designed to ensure safe deployment, continuous oversight, and controlled system evolution.

Preventive controls form the first defensive layer by reducing the probability that unsafe or non-compliant behaviour is triggered. Input sanitisation ensures that all data consumed by the agent — including prompts, documents, APIs, and data feeds — is free from malicious or misleading content, thereby protecting the ingestion stage. Prompt policies and approved templates standardise interactions and reduce behavioural variability while ensuring alignment with legal and compliance requirements. Guardrail libraries embed predefined behavioural and operational rules that prevent the agent from executing unsafe or policy-violating actions. Safe action space constraints further limit the agent’s operational perimeter by restricting access to systems, transactions, and irreversible commands. Additionally, human-in-the-loop checkpoints introduce expert oversight at critical decision points, particularly for high-impact activities such as financial recommendations or regulatory reporting.

While preventive measures reduce the likelihood of errors, **detective controls** identify emerging risks in real time. Logging and observability provide continuous traceability of agent behaviour, including tool usage, delegation chains, and deviations from expected workflows, thereby supporting auditability and early anomaly detection. Output validation systematically verifies agent-generated outputs against deterministic rules, statistical thresholds, or expert judgement, ensuring reliability and accuracy. Complementing these controls, explainability artefacts provide transparency through behavioural summaries and decision rationales, strengthening supervisory oversight and governance accountability.

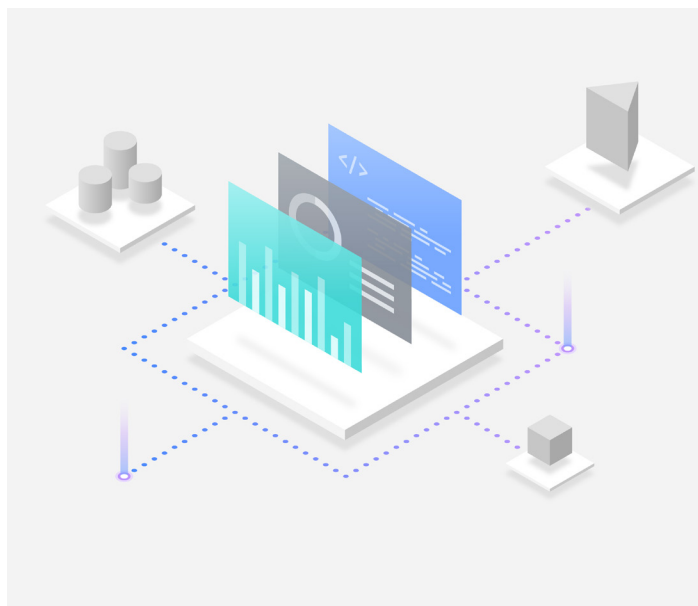
When anomalies or control breaches are detected, **adaptive controls** ensure effective response and long-term governance resilience. Independent change management ensures that any modification to the agent — including updates to model parameters, prompts, tools, or behavioural logic — undergoes formal risk assessment, validation, and approval. Structured incident management processes classify and escalate events based on model tier and impact severity, while lessons learned feed back into validation, monitoring calibration, and lifecycle governance. Automated containment mechanisms and interruptibility safeguards further enable organisations to halt or restrict agent behaviour when risk thresholds are exceeded.

From a 3LoD perspective, governance responsibilities remain clearly distributed. The First Line operates monitoring tools and executes containment actions. The Second Line defines escalation thresholds, analyses incident patterns, and evaluates residual risk against risk appetite. The Third Line independently assesses the effectiveness, timeliness, and governance escalation of monitoring and incident management frameworks.

6.4 Integration of Agentic AI Governance with Operational Risk Frameworks

In regulated financial institutions, the majority of agentic failures do not materialise as abstract model deficiencies, but as **operational events** affecting people, processes, systems, and external dependencies.

The combined framework therefore avoids treating Agentic AI as a purely technical or model-centric risk. Instead, it embeds agentic governance within standard operational risk structures, ensuring that incidents are identified, escalated, and remediated using mechanisms that are already supervisory-recognised and operationally mature. This integration strengthens operational resilience and prevents fragmentation between AI governance and enterprise risk management.



Operational Risk Domains Relevant to Agentic AI

Agentic failures map naturally to the four core operational risk domains commonly used across banking and insurance institutions, reflecting standard taxonomies:

- **People Risk:** Human actions or inactions impacting design, operation, oversight, or reliance on agentic systems, including inadequate training, overreliance on outputs, misconduct, or key-person dependencies.
- **Process Risk:** Failures in internal procedures, workflows, or controls, including uncontrolled delegation, broken validation checkpoints, or misalignment with internal policies.
- **Systems / Technology Risk:** Failures in IT infrastructure, software, data pipelines, or cybersecurity, including outages, software defects, data breaches, or attacks on agentic system environments.
- **External Events Risk:** Disruptions outside organizational control, such as cloud service outages, third-party model failures, geopolitical events, or systemic shocks affecting agentic operations.

By framing agentic failures through these established domains, the framework ensures that incidents are observable, classifiable, and manageable within existing ORM taxonomies, rather than being treated as exceptional or novel events.

Explicit Mapping Between MRM Risk Domains and Operational Risk

MRM defines **what must be controlled**, identifying critical agentic models, expected behaviour, and potential failure modes. ORM defines **how failures manifest**, using incident management, escalation, and remediation processes.

Each agentic-specific model risk domain naturally maps to one or more operational risk domains, ensuring that incidents are treated as standard operational events while remaining anchored in MRM’s forward-looking discipline:

- **MRM defines “what” must be governed** – identifying critical models, failure modes, and required preventive, detective, and adaptive controls.
- **ORM defines “how” failures manifest and are remediated** – leveraging incident management, escalation, and assurance mechanisms.
- **3LoD operationalise the integration:**
 - First line owns agentic execution and operational controls.
 - Second line aligns MRM and ORM frameworks.
 - Third line provides independent assurance across both model and operational risk dimensions.

This integration prevents fragmentation between AI governance and operational resilience, embedding Agentic AI within established risk structures and reinforcing supervisory credibility.

Table on Key Agentic AI Risks and Operational Risk Mapping

Agentic AI Risk	Primary Operational Risk Domain(s)
Hallucination / Fabrication	Process Risk; Systems/Technology Risk; People Risk
Autonomy Drift	Process Risk; Systems/Technology Risk
Delegation Escalation	Process Risk; People Risk
Reinforcement Learning Instabilities	Systems/Technology Risk; Process Risk
Cross-Entity Model Divergence	Process Risk; External Events Risk
Data Poisoning	Systems/Technology Risk; External Events Risk
Third-Party Dependence	External Events Risk; Systems/Technology Risk
Concentration Risk in Foundation Models	External Events Risk
Controllability and Interruptibility Risk	Systems/Technology Risk; Process Risk
Regulatory Non-Compliance Risk	Process Risk; People Risk

Integration of Explainability, Data, Governance, and Ethics across MRM, ORM, and 3LoD

Agentic AI Risk Domain	MRM Controls (Deviation & Benchmarking)	OR Controls (Prevention, Detection, Absorption)	3LoD – Ownership and Accountability
Explainability Risk	Definition of explainability standards; validation of interpretability relative to use-case materiality and benchmarks	Availability of explanations at decision, escalation, and customer interaction points; operational usability testing	1LoD applies explanations in practice; 2LoD challenges adequacy; 3LoD assures effectiveness
Data Risk	Assessment of data suitability, representativeness, sensitivity, and benchmark relevance	Data quality controls, lineage, access management, third-party data governance, resilience testing	1LoD manages data operations; 2LoD validates integrity; 3LoD audits end-to-end data governance
Governance Risk	Definition of decision rights, acceptable deviation ranges, escalation triggers	Embedding of governance rules into workflows, incident management, and recovery playbooks	1LoD executes governance; 2LoD enforces consistency; 3LoD provides assurance
Ethical / Conduct Risk	Evaluation of bias, fairness, and ethical trade-offs at model level using outcome-based testing	Human review, customer recourse, auditability, accountability mechanisms	1LoD applies ethical controls; 2LoD challenges outcomes; 3LoD assesses ethical governance over time

6.5 MRM and ORM as Complementary Control Frameworks for Agentic AI

MRM and ORM are [complementary, non-overlapping, and mutually reinforcing control frameworks for Agentic AI](#). Rather than treating Agentic AI as a new risk category requiring bespoke governance, this section demonstrates how existing, supervisory-recognised frameworks can be combined to address both [ex-ante model deviation and ex-post operational loss materialisation](#) in a coherent manner.

This section clarifies a critical distinction: MRM governs [why risk exists and how deviation from expected behaviour is assessed](#), while ORM governs [how failures are prevented, detected, absorbed, and remediated once systems operate in the real world](#). Neither framework is sufficient on a standalone basis for Agentic AI, given its autonomy, adaptive behaviour, and interaction with external tools and environments.

This positioning is fully consistent with prudential expectations across banking and insurance supervision, including Basel model risk principles, IAIS and Solvency II governance requirements, OSFI Guideline E-23, and supervisory thinking on operational resilience in the United States and Europe.

Allocation of responsibilities across MRM and ORM

MRM remains anchored in [benchmarking, deviation, and conceptual soundness](#). In the context of Agentic AI, this includes assessing deviation from market benchmarks, policy benchmarks, or human-performed reference activities, depending on the nature of the task. Structural deviation — not statistical noise — is treated as model risk.

ORM, by contrast, addresses [loss scenarios, failure frequencies, and impact](#), including those arising from malfunctioning agents, inadequate training, misconfiguration, failed human-in-the-loop controls, or external disruptions. Importantly, ORM is not reduced to incident management: preventive controls are explicitly recognised as part of operational risk governance.

The two frameworks interact structurally: MRM informs ORM scenarios by defining where deviation is possible and material, while ORM validates whether MRM-defined constraints and controls are effective in practice.

Integration of explainability, data, governance, and ethics across MRM and ORM

Explainability, data quality, governance, and ethical outcomes are treated as [shared control dimensions](#), not as standalone risk silos or parallel “responsible AI” frameworks.

Within MRM, these dimensions are assessed at model level: explainability requirements are defined proportionately; data suitability and benchmark relevance are validated; acceptable deviation ranges and decision rights are specified; and ethical risks are evaluated as outcome-oriented model risks.

ORM ensures that these same dimensions are [operationally effective](#): explanations are usable at decision and escalation points; data controls cover lineage, access, third-party dependencies, and resilience; governance rules are embedded into workflows and incident management; and ethical intent is enforced through human review, auditability, and customer recourse mechanisms.

This approach embeds widely recognised international principles – often articulated under headings such as explainability, data governance, accountability, and ethical outcomes – [without creating a parallel framework](#) or duplicating ownership.

Distributed accountability across the 3LoD

Accountability for Agentic AI risk is deliberately distributed across the 3LoD.

The first line owns agentic outcomes and executes both model-level and operational controls. The second line independently challenges model deviation, data integrity, ethical outcomes, and operational effectiveness against risk appetite and supervisory expectations. The third line provides assurance that the integrated framework functions as intended over time, including under conditions of system adaptation and change.

Human-in-the-loop controls are explicitly framed as [designed risk mitigants](#), not emergency fallbacks. In high-stakes and regulated contexts, residual error rates that may be acceptable elsewhere constitute structural model risk. Human judgment therefore remains a mandatory control for Tier 1 and Tier 2 Agentic AI systems, with clearly defined authority, documentation standards, and accountability.

The table above operationalises the core thesis of the white paper by making explicit the [end-to-end chain of accountability](#) for Agentic AI risk.

The table clarifies:

- [What can go wrong](#) through the definition of Agentic AI risk domains (MRM lens);
- [Why the risk exists and how deviation is assessed](#) through MRM controls anchored in benchmarking;
- [How failures are prevented, detected, and absorbed](#) through ORM controls that operate before and after deployment;
- [Who is accountable](#) through a clear allocation across the 3LoD.

By presenting these dimensions side by side, the table demonstrates that:

- MRM defines acceptable behaviour and deviation;
- ORM ensures resilience and loss prevention through explicit preventive and reactive controls;
- the 3LoD make both effective, auditable, and supervisory-credible.

95%

Accuracy levels that may be acceptable in low-impact or consumer contexts but are insufficient in high-stakes and regulated environments due to the material consequences of residual error.

6.6 Key Takeaways of the Holistic MRM, ORM and 3LoD Framework

Agentic AI can be governed end-to-end using the same disciplined, risk-based logic applied to other high-impact models, while explicitly addressing the additional risks introduced by autonomy, behavioural complexity, and real-world action. Taken together, these sections clarify how the combined framework answers the fundamental governance questions raised by Agentic AI:

- **Tiering** establishes how critical the model is from a risk perspective, based on materiality, model usage, and external impact.
- **Lifecycle redesign** explains how that criticality is governed over time as the system operates, adapts, and evolves.
- **Validation extensions** define how risk is independently assessed and challenged in the presence of autonomous and adaptive behaviour.
- **Monitoring and incident response** determine how loss events and control breaches are detected, contained, and prevented once the model is live.

This is precisely where the innovation of the combined framework lies. **MRM** provides the structural backbone that translates model criticality into concrete, tier-driven control expectations, while the **3LoD** ensure that these expectations are owned by the business, independently challenged by risk functions, and subject to ongoing assurance through audit.

Together, MRM and the 3LoD enable Agentic AI to be governed in a coherent and supervisory-credible manner — without creating parallel, AI-specific control regimes and while remaining fully aligned with established governance and supervisory principles.

Importantly, the framework also makes explicit the **operational risk dimension** of Agentic AI. While Agentic AI is governed through MRM, many of its failure modes — such as erroneous execution, control bypass, unsafe tool use, or cascading multi-agent errors — materialise operationally rather than statistically. By embedding monitoring, incident management, escalation, and change controls within the same tier-driven governance structure, **the framework ensures that Agentic AI incidents are treated as first-class operational risk events, fully integrated into existing operational risk and incident management processes.** This alignment prevents gaps between model governance and operational resilience and reinforces the effectiveness of the 3LoD in practice.

Last but not least, in **high-stakes and regulated environments, performance levels that may be acceptable in consumer or low-impact contexts, for example, 90–95% correctness, are insufficient.** From a regulatory and risk perspective, such residual error rates constitute **structural model risk, not statistical noise.** As a result, **human-in-the-loop controls are not optional safeguards but mandatory risk mitigants** for Tier 1 and Tier 2 Agentic AI systems.

Crucially, human oversight within this framework is not limited to prompt supervision or exception handling. The human remains accountable for the final validation of outputs, particularly where agentic systems generate regulatory artefacts, financial calculations, risk assessments, or decisions with legal or prudential consequences. In this sense, human judgment functions as a critical control within both the model risk and operational risk domains, ensuring that residual automation risk is consciously managed rather than implicitly accepted.

Enterprise Use Cases: From Governance Principles to Operational Value

The preceding sections have established that Agentic AI can be governed effectively when it is embedded within MRM and fully integrated into the 3LoD. This section demonstrates how this combined framework translates into **concrete business value**, while preserving **regulatory credibility and control**, through two representative enterprise use cases.

The cases are intentionally contrasted. In the insurance example, Risk Management acts primarily as a **user** of Agentic AI to enhance analytical capability and decision support. In the banking example, Risk Management acts as a **steward**, governing and challenging agentic systems deployed within the business. Together, they illustrate the flexibility of the framework and its applicability across sectors, roles, and organisational configurations.

7.1 Insurance Use Case: Risk Management as User of Agentic AI

Context and Business Objective

In the insurance sector, Risk Management functions are increasingly required to support complex, forward-looking assessments, including Own Risk and Solvency Assessment (ORSA), capital projections, reserving analysis, and sensitivity testing under multiple economic and demographic scenarios.

These activities are analytically intensive, time-constrained, and heavily reliant on expert judgement.

Agentic AI offers a clear business advantage in this context. By orchestrating scenario generation, model execution, narrative interpretation, and consistency checks, an agentic system can significantly enhance the **speed, breadth, and coherence** of risk analysis. The value proposition is not automation for its own sake, but improved **decision support quality**, particularly under stress conditions where management needs timely and structured insight.

Role of Risk Management and Tiering Implications

In this use case, Risk Management acts as the **first-line user** of the agentic system. The agent supports risk analysts by proposing scenarios, interpreting model outputs, and drafting analytical narratives. While the agent does not execute financial transactions or directly affect policyholder outcomes, its outputs influence internal risk assessments and management decisions.

Materiality is elevated because outputs inform capital and solvency analysis; usage is high due to repeated and embedded analytical support; and external impact remains indirect but relevant, particularly where ORSA outcomes are reviewed by supervisors.

Tiering is therefore not a formality. It determines the depth of validation, documentation, and oversight required, ensuring that the analytical power of the agent does not come at the expense of governance discipline.

Embedded Controls and Combined Governance

The core controls introduced in Section 6.3 are embedded directly into the lifecycle and operation of the agentic system. Preventive controls play a central role. Input sanitisation ensures that economic assumptions, stress narratives, and data feeds used by the agent are free from ambiguous constraints or biased inputs. Prompt policies and approved prompt templates standardise how analytical questions are framed, reducing variability and ensuring alignment with internal risk methodologies.

Output validation operates as a detective control, checking generated analyses, capital impacts, and narrative explanations against tolerance thresholds and expert expectations. For higher-impact outputs, human-in-the-loop checkpoints ensure that a qualified risk professional reviews and confirms conclusions before they are escalated to senior management.

Logging and observability provide continuous traceability of how scenarios were generated, which tools were invoked, and how conclusions were derived. Explainability artefacts support transparency by enabling reviewers—internal or supervisory—to understand not only the results, but the reasoning path followed by the agent.

From a 3LoD perspective, accountability is clear and coherent. The first line (Risk Management) operates the agent and applies controls in practice. The second line defines standards for acceptable use, validation depth, and documentation, and independently challenges the reliability of outputs. The third line later assesses whether the use of agentic AI within Risk Management remains consistent with the institution's risk appetite and governance expectations.

Business and Regulatory Benefits

This use case illustrates how the framework enables **measurable business benefits**—faster analysis, richer scenario coverage, and more consistent narratives—without diluting control. At the same time, governance remains aligned with insurance supervisory expectations on model governance, transparency, and ORSA robustness. Agentic AI enhances Risk Management capability while remaining fully embedded within existing MRM and 3LoD structures.

7.2 Banking Use Case: Risk Management as Steward of Agentic AI

Context and Business Objective

In banking, agentic AI is increasingly deployed within the **business lines** rather than within Risk Management itself. Typical use cases include credit assessment support, portfolio monitoring, pricing optimisation, liquidity forecasting, and client interaction tools. These systems often operate close to execution and may influence financial outcomes, customer treatment, and regulatory metrics.

In this environment, Risk Management's primary role is not to use the agent, but to **govern, challenge, and oversee it**. The business objective is to enable innovation and efficiency within the first line, while ensuring that autonomy, complexity, and third-party dependence do not undermine prudential soundness or supervisory trust.

Stewardship Role and Tiering Consequences

Here, Risk Management acts as the **steward** of agentic AI within the MRM framework. Applying the tiering methodology, many banking agentic systems naturally fall into **Tier 1** or **Tier 2**, particularly where they influence credit decisions, capital metrics, or pricing outcomes. High model usage, direct external impact, and interaction with customers or supervisors drive tier escalation.

Tiering becomes the anchor through which stewardship is exercised. It determines not only validation intensity, but also the scope of lifecycle controls, monitoring requirements, and escalation pathways. Crucially, it provides a regulator-credible basis for challenging business proposals and constraining autonomy where risk appetite would otherwise be exceeded.

Holistic Controls Across the Lifecycle

The core controls described in Section 6.3 are deployed across the agent's lifecycle as an integrated control fabric. Preventive controls include safe action space constraints that restrict what the agent can execute, particularly in relation to credit approvals, pricing adjustments, or system access. Guardrail libraries encode policy and regulatory constraints directly into agent behaviour, reducing reliance on ex-post detection.

Detective controls play a critical role in a live banking environment. Logging and observability enable continuous monitoring of tool use, delegation chains, and multi-agent

interactions. Output validation and behavioural monitoring identify deviations from expected patterns, supporting early intervention before customer harm or prudential breaches occur.

Adaptive controls ensure that governance remains effective as systems evolve. Independent change management requires that any modification to prompts, tools, or model components be assessed and approved in line with MRM change standards. This prevents incremental erosion of controls as agentic capabilities expand.

From a 3LoD perspective, the business owns and operates the agentic system, embedding controls into daily processes. Risk Management defines standards, performs independent validation and challenge, and escalates issues where tier-based expectations are not met. Internal Audit provides assurance that stewardship is effective, particularly for high-tier systems with material financial or supervisory impact.

Business Enablement and Supervisory Credibility

This use case demonstrates that strong governance does not inhibit innovation. On the contrary, by providing clear, tier-based expectations and embedded controls, the framework enables the business to deploy agentic AI with confidence. Risk Management's stewardship role ensures consistency, comparability, and auditability across business lines and legal entities.

From a regulatory perspective, this approach aligns with supervisory expectations under CRR/CRD, SR 11-7-aligned practices, and operational resilience regimes. Agentic AI is governed through existing review channels — internal model approvals, thematic inspections, and horizontal governance reviews — rather than through ad hoc or AI-specific processes.

7.3 Synthesis: A Unified Framework Across Roles and Sectors

Taken together, the two use cases demonstrate the versatility and robustness of the combined Model Risk Management and the Three Lines of Defence framework. Whether Risk Management acts as a [user](#) or a [steward](#), the same underlying principles apply: tiering determines criticality; core controls operationalise governance; and accountability is distributed clearly across the three lines.

The result is a framework that delivers [business advantage](#), [holistic control](#), and [regulatory compliance](#) simultaneously. Agentic AI becomes a managed source of value rather than an unmanaged source of risk, fully embedded within the institution's existing governance architecture rather than isolated at its margins.

Key Takeaways

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 Model Risk Management and Operational Risk within the Three Lines of Defence](#)
4. [Translation into Practice: Lessons from the Two Use Cases](#)
5. [Regulatory Expectations, Best Practices, and Cross-Jurisdictional Gaps](#)

8.1 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.

8.2 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 white paper 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 white paper 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.

8.3 Innovation in Controls: Preventive, Detective, and Adaptive Governance at the Intersection of MRM and Operational Risk 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. They [include 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.

8.4 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, Operational Risk, and 3LoD approach clarifies ownership despite increased autonomy, enables independent challenge without stifling innovation, and provides a defensible narrative for supervisors across jurisdictions.

8.5 Regulatory Expectations, Best Practices, and Cross-Jurisdictional Gaps

A final contribution of the white paper lies in its cross-jurisdictional regulatory analysis, consolidated in the Annex. Through a structured comparison of regulatory obligations, industry best practices, and regulatory gaps across the EU, UK, US, and Canada, the white paper 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.

This document is complemented by a handbook¹⁹, 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 Lascale
Promontory Director
olascale@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

References

- [1] AI Act, “Regulation (EU) 2024/1689 del Parlamento Europeo e del Consiglio”, 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=OJ:L_202401689
- [2] BIS, “The use of artificial intelligence for policy purposes”. 2025. [Online]. Available: <https://www.bis.org/publ/othp100.pdf>
- [3] BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM, “SR 11-7: Guidance on MRM (Supervision and Regulation Letters)”, 2011. [Online]. Available: <https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf>
- [4] Commission Delegated Regulation (EU), “2015/35 of 10 October 2014 supplementing Directive 2009/138/EC of the European Parliament and of the Council on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II)”. 2014. [Online]. Available: Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II)
- [5] Cornell University, “Levels of Autonomy for AI Agents”, 2025. [Online]. Available: [2506.12469] Levels of Autonomy for AI Agents
- [6] Cornell University, “TRISM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems”, 2025. [Online]. Available: [2506.04133] TRISM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems
- [7] EIOPA, «Opinion on Artificial Intelligence governance and risk management.pdf», 2025. [Online]. Available: https://www.eiopa.europa.eu/publications/opinion-artificial-intelligence-governance-and-risk-management_en
- [8] European Parliament, “REPORT on the impact of artificial intelligence on the financial sector”, 2025. [Online]. Available: REPORT on the impact of artificial intelligence on the financial sector | A10-0225/2025 | European Parliament
- [9] FSB, “The Financial Stability Implications of Artificial Intelligence”. 2024. [Online]. Available: <https://www.fsb.org/uploads/P14112024.pdf>
- [10] Global Risk Institute, “A Canadian Perspective on Responsible AI”, 2025. [Online]. Available: Financial Industry Forum on Artificial Intelligence: A Canadian Perspective on Responsible AI - Global Risk Institute
- [11] IAIS (International Association of Insurance Supervisors), “Application Paper on the supervision of artificial intelligence”. 2025. [Online]. Available: <https://www.iais.org/uploads/2025/07/Application-Paper-on-the-supervision-of-artificial-intelligence.pdf>
- [12] IAIS (International Association of Insurance Supervisors), “Regulation and supervision of AI/ML – a thematic review”. 2023. [Online]. Available: <https://www.iais.org/uploads/2023/12/Regulation-and-supervision-of-AI-ML-a-thematic-review.pdf>
- [13] IBM, “IBM Whitepaper: Accountability and Risk Matter in Agentic AI”, 2025. [Online]. Available: IBM Whitepaper: Accountability and Risk Matter in Agentic AI
- [14] NAIC, “Model Laws, Regulations, Guidelines and Other Resources - ANNUAL FINANCIAL REPORTING MODEL REGULATION”. 2015. [Online]. Available: <https://content.naic.org/sites/default/files/model-law-205.pdf>
- [15] NATURE, « Delegation to artificial intelligence can increase dishonest behavior », 2025. [Online]. Available: Delegation to artificial intelligence can increase dishonest behavior | Nature
- [16] Okpala, A. Golgoon, and A. R. Kannan, “Agentic AI Systems Applied to Tasks in Financial Services: Modeling and MRM Crews, arXiv:2502.05439v2 [cs.AI]”, 2025. [Online]. Available: <https://arxiv.org/abs/2502.05439>
- [17] OSFI, “Guideline E-23 – MRM 11/2025”. 2025. [Online]. Available: <https://www.osfi-bsif.gc.ca/en/guidance/guidance-library/guideline-e-23-model-risk-management-2027>
- [18] PRA, “SS1/23 – MRM principles for banks - Supervisory statement 1/23”. 2023. [Online]. Available: <https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/supervisory-statement/2023/ss123.pdf>
- [19] Promontory, “A Practitioner Handbook”. M. Onorato, O. Lascala, S. Peschiera, S. Procacci, G. Ugolini, R. Cinelli, M. Muscolo. 2026. [Online].
- [20] US Department of the Treasury, “Artificial Intelligence in the Financial Services”, 2024. [Online]. Available: Artificial-Intelligence-in-Financial-Services.pdf
- [21] University of Chicago, “The Law of AI is the Law of Risky Agents Without Intentions | The University of Chicago Law Review”, 2025. [Online]. Available: The Law of AI is the Law of Risky Agents Without Intentions | The University of Chicago Law Review

© Copyright IBM Promontory 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.

