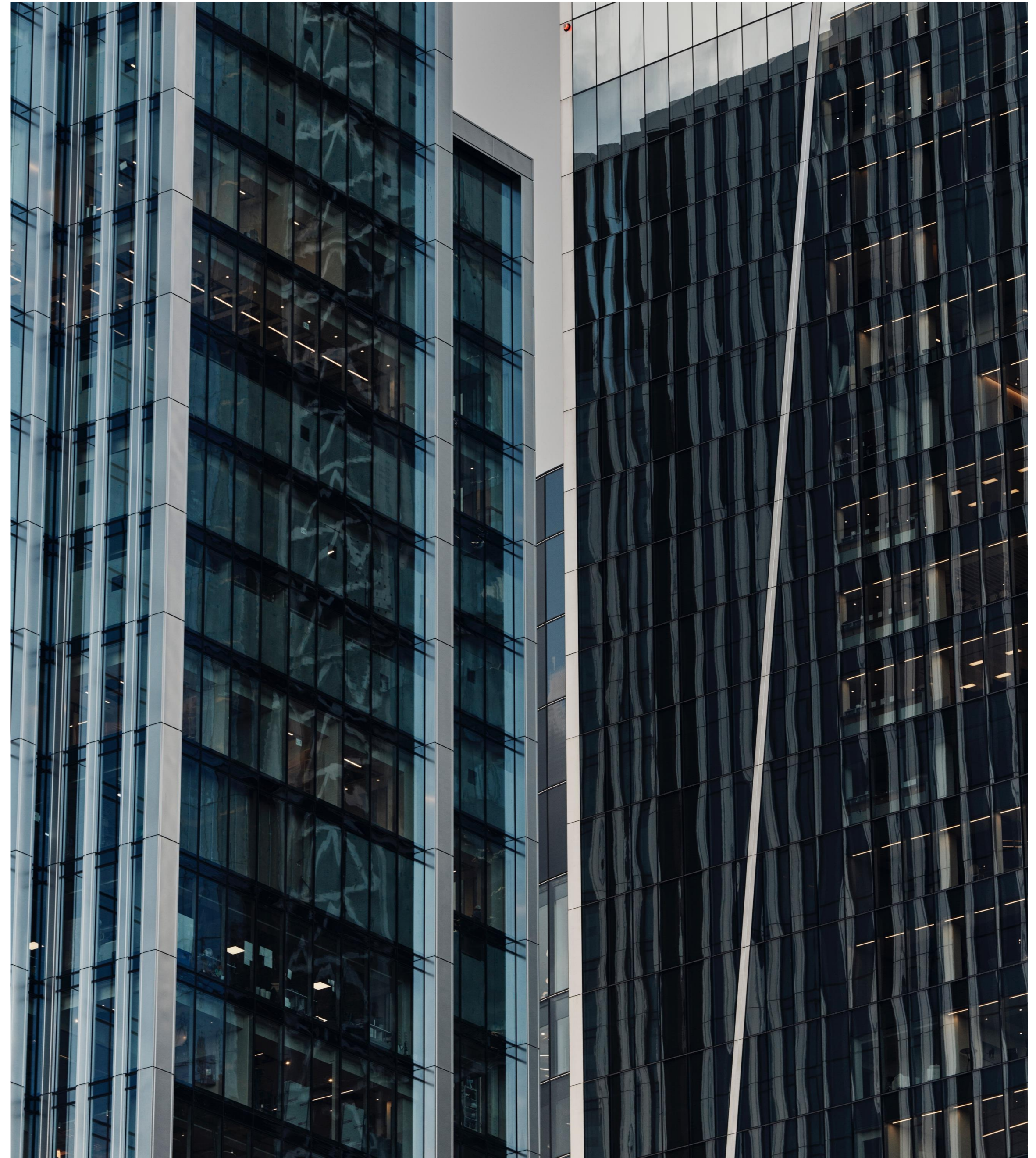


Governing Agentic AI in Financial Institutions

Integrating Model Risk Management with the Three Lines of Defense into a Holistic Control Framework for Regulatory and Best Practice Compliance: Practical Use Cases for Risk Managers.



[Full Document Repository](#)



How can financial institutions scale Agentic AI while preserving control, accountability, and supervisory credibility?

Governing Agentic AI in financial institutions does not require a new, AI-specific governance paradigm. Because agents plan and act through autonomy, tool use, delegation, and dynamic execution, risk becomes behavioral and continuous at runtime, so governance must be continuous too. The proposed approach embeds agentic systems into existing [Model Risk Management](#) (tiering, lifecycle, validation, monitoring, change) and allocates clear responsibilities across the Three Lines of Defense.

Done this way, autonomy can be scaled while preserving [accountability](#), [auditability](#), and [supervisory credibility](#) across jurisdictions.

Key Takeaways

Scale with Control

Unlock agentic productivity while keeping decision ownership, auditability, and risk within bounds

Static Governance Fails

Agent behavior evolves at runtime, and pre-launch reviews miss delegation, tool use, compounding actions

Adaptive Controls

Embed preventive, detective, adaptive controls in MRM and Three Lines of Defense for continuous assurance

Use-Case Lessons

Apply one framework across insurance and banking, and clarify Risk Management as User and Steward

Regulatory Readiness

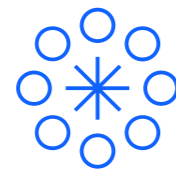
Meet converging expectations and manage gaps across jurisdictions with evidence, monitoring, and accountability

The combined Model Risk Management and Three Lines of Defense framework enables institutions to **scale agentic AI for faster, more adaptive decisions** while preserving **accountability, auditability, and ownership** through controlled autonomy that is bounded, benchmarked, observable, and enforceable.

-
- Agentic AI accelerates analysis and enables continuous, adaptive decision support across banking and insurance.
 - Without a strong governance backbone, it amplifies operational, conduct, and prudential risks (error propagation, weaker oversight, more dependencies).
 - Business value comes from controlled autonomy: bounded behavior, measurable benchmarks, observability, and clear accountability.
 - **Insurance (RM as user):** better scenario/capital insights while preserving actuarial judgement and management ownership.
 - **Banking (RM as steward):** safe use in capital models, stress testing, and aggregation without undermining integrity, resilience, or supervisory confidence.
 - Overall, benefits arise when agentic AI is embedded as a risk-bearing system within core governance and operational controls.

Agentic AI changes how risk propagates in production, meaning that static approvals and periodic reviews are insufficient: governance **must integrate** model risk and operational risk disciplines.

Dynamic Risk



Autonomy, multi-step reasoning, tool use, and delegation create non-linear behaviors that evolve during execution, not just at design time, changing how errors and impacts propagate across processes.

Beyond Approval



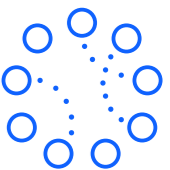
Initial approval, documentation, and periodic review remain necessary, but risk does not stabilize at sign-off; it evolves as agents interact with data, tools, users, and external dependencies, and can escalate to higher tiers.

Use-case Impact



Even “advisory” agents can influence capital and reserving decisions in insurance, while embedded overlays in banking amplify downstream impact and supervisory expectations, even if formal model classification stays unchanged.

MRM + ORM Synergy



Governance must treat agentic systems as both model risk (structural deviation from benchmarks/challengers/expert judgement) and operational risk (execution failures, controls, data, HITL, misconfiguration, third parties), linking acceptable behavior to resilient operations.

Innovation comes from embedding **preventive**, **detective**, and **adaptive controls** into existing MRM, anchored by ORM and executed through the 3LoD AI principles into auditable, enforceable requirements without creating parallel governance.

Governance Control Layers

Preventive Controls

Upfront guardrails that bound behavior (prompt policies, action limits, guardrails, HITL), supported by access/SoD and resilience.

Detective Controls

Ongoing monitoring to catch divergence, hallucinations, drift, or unauthorized delegation (validation, logs, observability) and trigger escalation.

Adaptive Controls

Change and incident controls to stay effective over time (change management, escalation, response) across updates and scope changes.

Responsible-AI Risk Drivers

Explainability Risk

Decisions aren't traceable or challengeable; addressed via provenance, logging, and tiered validation/escalation.

Data Risk

Poor quality/lineage/timeliness amplified by autonomy and tools; managed via data suitability tests plus data governance and access controls.

Ethical Risks

Bias/fairness harms emerge over time; controlled through measurable constraints, testing/monitoring, and accountability/escalation.

The two use cases show how a single MRM–Operational Risk–Three Lines of Defense **governance framework** can flex across insurance and banking while preserving a common control logic, clarifying ownership despite autonomy, enabling independent challenge **without** stifling innovation, and supporting a defensible, cross-jurisdiction supervisory narrative.

Common Control Logic

- 1. Role Clarity** - Define who owns decisions (user vs steward) and what remains human judgement.
- 2. Bounded Autonomy** - Specify permitted actions, tool access, and hard stop / HITL checkpoints
- 3. Evidence Pack** - Document prompts, data inputs, tool calls, outputs, and decision provenance for review/audit
- 4. Ongoing Assurance** - Monitor behavior and deviations with tier-proportionate thresholds and reporting.
- 5. Operational Anchoring** - Route issues through existing incident, escalation, and change-management processes.

Agentic AI is materially governed across jurisdictions through existing prudential, conduct, governance, and operational resilience regimes: with no explicit autonomy rulebook, institutions **must meet convergent supervisory outcomes** using robust internal MRM–ORM–3LoD controls.

Regulatory Convergence

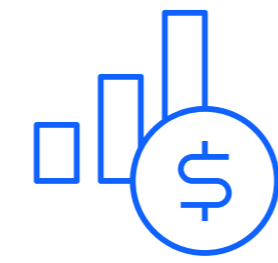
Agentic AI is already governed, often indirectly, through existing prudential, conduct, governance, and operational resilience frameworks.

While no jurisdiction offers a complete rulebook for agentic autonomy, **supervisors converge on outcomes**: accountability, controllability, explainability, proportionality, resilience, and auditability. Where guidance is silent (multi-agent behavior, dynamic delegation, behavioral drift), institutions **are expected** to rely on robust internal governance and operational risk disciplines.

The combined MRM and Three Lines of Defense framework therefore acts as a practical regulatory and resilience strategy, translating best practices into auditable controls and enabling credible supervisory engagement across jurisdictions.

No major jurisdiction regulates “Agentic AI” as a standalone category, but through **existing legal and prudential regimes.**

Sectoral Overview



Banking Sector

Agentic AI in banking is increasingly used in high-impact areas (credit decisioning, capital planning, stress testing, pricing, liquidity), and is governed through a mix of regimes:

- EU layers AI Act obligations onto internal model governance
- UK relies on PRA/FCA expectations (model risk, accountability, operational resilience)
- US applies SR 11-7 and interagency MRM guidance (only partially covering agentic behavior)
- Canada uses OSFI E-23 (strong on governance/resilience, limited on agent-level controls)



Insurance Sector

In insurance, agentic AI is emerging in actuarial work, underwriting, claims, and fraud. Governance expectations converge on model risk discipline, policyholder protection, and supervisory transparency:

- Solvency II is robust but not designed for autonomous agents
- UK and North America are mainly principles-based (no formal tiering)
- Supervisors increasingly expect explainability, auditability, and control, even when models are not formally “regulatory”

Risk Management
functions are also
users and beneficiaries
of Agentic AI

This Duality Introduces a Structural Tension

Risk Management
must **rely on**
agentic systems
while simultaneously
challenging and
governing them

Proposed Approaches

Renewal of Policies & Standards

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

Focus of Risk Management

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

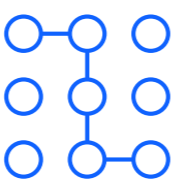
Existing tiering frameworks remain applicable, with Agentic AI **amplifying** traditional model risk dimensions.



Materiality

Materiality rises when agent outputs influence financial results, capital/reserves, pricing, or client outcomes at scale.

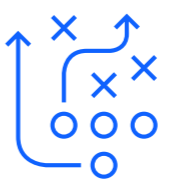
In agentic workflows, impact is often indirect and cumulative: agents break objectives into sub-tasks, use tools, exchange intermediate outputs.



Model Usage

Embedding agents leads to increments in model usage, and often operate continuously, pulling from multiple sources and triggering downstream actions via tools/APIs.

In multi-agent architectures, usage expands through dynamic delegation, so reliance and scope can evolve at runtime



External Impact

External impact expands when agent outputs affect customers, counterparties, supervisors, or public disclosures. In federated deployments, agents may be tuned or used differently across entities, creating cross-entity divergence and undermining comparability and auditability. Dependence on third-party models/APIs and external data pipelines further increases exposure

AI escalates quickly because runtime behavior and autonomy create compounding, expanding risk → Tier 1 & 2 classification must trigger enforceable governance outcomes.

Risk domains that drive Tier Escalation

Hallucination / Fabrication	Incorrect outputs can compound across multi-step tasks and tool calls
Autonomy Drift	Behavior gradually deviates from intended boundaries during live operation
Delegation Escalation	Agents delegate beyond authorised scope, expanding the action perimeter
Cross-entity Divergence	Same agent behaves differently across entities, reducing comparability/auditability
Third-Party Dependence	Reliance on external models/APIs increases dependency risk and reduces controllability

Tiering Classification

Validation intensity	Deeper / extended validation (including behavioural stability), more frequent reviews
Lifecycle Controls	Mandatory preventive / detective / adaptive controls embedded end-to-end
Monitoring Requirements	Continuous monitoring / logging (including tool use and delegation chains)
Escalation Pathways	Defined thresholds, governance gates, and escalation routes aligned to tier
Change Governance	Independent change management for prompts / tools / components to prevent control erosion

Integration of the
Model Risk
Management with the
Three Lines of Defence
into a single,
enforceable control
system for Agentic AI

1°

The **first line** owns and operates agentic systems

2°

The **second line** defines standards and performs independent challenge

3°

The **third line** provides assurance on effectiveness

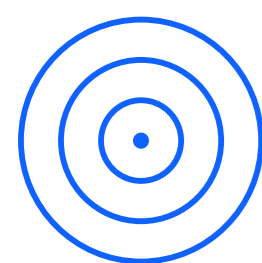
Explicit risk domains **bridge** abstract tiering criteria and concrete controls.

Tiering classifies agentic criticality, but it is not executable on its own: explicit risk domains translate tiering outcomes into concrete control requirements (what must be constrained, tested, monitored, and evidenced) so governance moves from classification to enforceable action.

Model Risk Management provides the forward-looking discipline to govern autonomy, behavioural deviation, and emergence. Governance becomes effective only when MRM is integrated with Operational Risk processes (resilience, change, incidents) and embedded into existing accountability structures.

This integration avoids AI-specific silos while remaining supervisory-credible. It applies consistently across banking and insurance: high-impact agentic use cases are governed with the rigour of other material models, while explicitly recognising their additional operational risk footprint—preventing risk ex ante and managing it when it materialises.

Agentic AI controls are embedded across lifecycle, validation, monitoring, and change management: **Tiering is the primary control lever**, as it determines control intensity and assurance, enforced through the Three Lines of Defense to keep agentic risk continuously controlled.



Preventive

Tiering sets a safe action space that limits tools, permissions, data access, delegation, and execution rights.

Inputs and prompts are sanitised and standardised through approved policies and templates, with safeguards for sensitive data.

Validation scales by tier, with human checkpoints for higher tiers.



Detective

Once running, outputs are validated against rules and benchmarks, while logging and observability make every step traceable, that is tools used, data accessed, delegation chains, and decision steps.

Continuous monitoring flags drift, hallucination, or scope creep and triggers predefined escalation paths, including revalidation or constraint tightening.



Adaptive

Independent change governance controls modifications to prompts, tools, integrations, models, and operating scope so risk does not creep in silently.

Incidents trigger containment and remediation, and the learning loop feeds back into control redesign.

Material changes or events can require recertification and tier reassessment.

Integration of Agentic AI Governance with Operational Risk Frameworks

Agentic AI governance, anchored in Model Risk Management, must **explicitly intersect** with established Operational Risk frameworks to capture real-world risk materialisation.

Operational Risk Domains Relevant to Agentic AI

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

What

MRM defines “what” must be governed – identifying critical models, failure modes, and required preventive, detective, and adaptive controls.

How

Operational Risk defines “how” failures manifest and are remediated – leveraging incident management, escalation, and assurance mechanisms.

Who

Three Lines of Defence operationalise the integration:

- First line owns agentic execution and operational controls
- Second line aligns MRM and operational risk frameworks.
- Third line provides independent assurance across both model and operational risk dimensions.

Model Risk Management (MRM) and Operational Risk Management (ORM) are **complementary**, **non-overlapping**, and **mutually reinforcing** control frameworks for Agentic AI

Risk domains that drive Tier Escalation

MRM governs why risk exists by assessing deviation from expected behaviour.

For agents, it focuses on conceptual soundness, benchmark selection, acceptable deviation ranges, autonomy boundaries, and revalidation triggers. Structural deviation—not statistical noise—defines model risk and drives assurance and documentation.

Tiering Classification

ORM governs how failures are prevented, detected, absorbed, and remediated in live operations.

It covers loss scenarios and resilience drivers such as tool outages, third-party dependencies, misconfiguration, failed human-in-the-loop controls, process breakdowns, and incident response and recovery playbooks.

Together

MRM identifies where material deviation is possible and defines acceptable boundaries, while ORM verifies that constraints work in real operations and closes the loop through incident response, controlled change management, and tier reassessment

Agentic AI can be governed end-to-end using established risk-based practices, extended for autonomy and runtime behaviour.

MRM converts tiering into tier-driven controls, and the Three Lines of Defence makes them enforceable

Because failures often materialise as operational incidents, monitoring, incident management, and change controls must be integrated in the same tier-driven structure as to treat Agentic AI incidents as **first-class operational risk events**, fully integrated into existing operational risk and incident management processes.

Tiering

Defines criticality (materiality, usage, external impact)

Lifecycle

Governs criticality over time as systems operate and evolve

Validation

Extends independent challenge to autonomous / adaptive behaviour

Monitoring and Response

Detects, contains, and prevents live control breaches

90-95%

performance level in consumer and low impact systems that is however not acceptable in high-stakes environments as it constitutes structural model risk

Human-in-the-loop controls are not optional safeguards but mandatory risk mitigants for Tier 1 and Tier 2 Agentic AI systems

A domain-specific Documentation Crew [synthesises internal technical evidence](#) into regulator-ready ORSA Risk Profile text under human supervision and without changing existing governance processes.

Business Problem

ORSA is a core Pillar II deliverable (Solvency II / NAIC-style regimes): supervisors assess credibility, coherence, and governance integration, not only numbers.

The Risk Profile section requires an enterprise-wide narrative across risks and interactions, backed by evidence.

Scope & Outputs

A structured draft for SCR Breakdown subsections, at minimum:

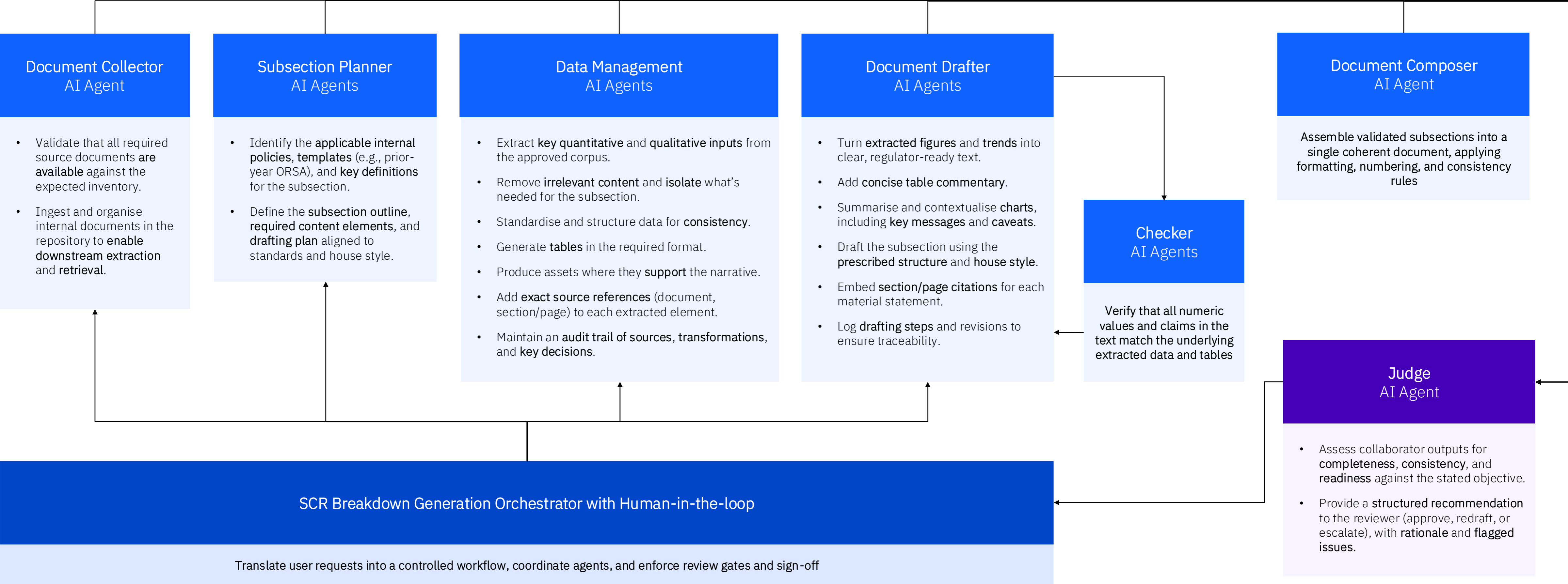
- Underwriting Risk
- Market Risk
- Credit Risk
- Liquidity Risk
- Operational Risk
- Other Material Risks

Output is [house-style compliant](#), with embedded citations to source docs and optional tables / charts placeholders.

Key Outcomes

- Accelerate ORSA drafting and review through a controlled agentic workflow, reducing cycle time and manual effort.
- Standardise **annual ORSA outputs** with consistent structure and templates, improving comparability across cycles.
- Provide **source-grounded, audit-ready evidence** so SMEs focus on challenge and approval, not synthesis.

A Human-in-the-loop Documentation Crew [orchestrates](#) specialised agents to extract evidence, draft regulator-ready SCR Breakdown text, and enforce validation and auditability through [checkers](#), [citations](#), and [structured review](#)



Risk Management operationalises second-line stewardship by [independently tiering](#), [challenging](#), and [approving](#) business-owned agentic AI under MRM + 3LoD, [producing auditable evidence](#) for supervisors and internal assurance.

1st LoD

Finance / Business

Designs and operates the solution, embeds safeguards, and provides evidence on controls and performance

2nd LoD

Risk Management

[Sets validation expectations, performs independent review and challenge, assigns tiering, and decides approve / conditional approve / escalate based on residual risk](#)

3rd LoD

Audit

Later assures that 1LoD and 2LoD activities were effective, independent, and consistently applied

6-step Stewardship [Review Workflow](#)

A tier-driven, evidence-based review that assesses agent autonomy and behaviour against MRM requirements, documents controls and residual risk, and results in a clear outcome: [approve](#), [conditional approval](#), or [escalation](#).

1

Prompt & Objective Setting
Confirm the objective is bounded and unambiguous, prompts are approved/versioned, and that autonomy level and use-case scope match risk appetite.

2

Data Aggregation Rules & Tool-Use Boundaries
Verify authorised sources, lineage, and reproducibility, enforce data minimisation and access controls, and confirm tool permissions, limits, and logging.

3

Relationship Mapping & Risk-Flagging Logic
Assess how the system infers relationships and triggers flags, ensure traceability/explainability, test accuracy and false positives/negatives, and check consistency vs internal taxonomies.

4

Human-in-the-loop (HITL) & Refinement Controls
Validate mandatory approval gates for high-impact actions, ensure override/interruptibility and escalation paths, and prevent uncontrolled learning or autonomy expansion.

5

Narrative Generation & Source Attribution
Ensure outputs are grounded in verifiable inputs, require per-claim citations, apply hallucination and consistency checks, and confirm disclosure controls for sensitive information.

6

Scenario & Stress Testing
Design and run adverse/edge-case scenarios (conflicting objectives, degraded inputs, tool failures, adversarial prompts), test delegation chains where relevant, and document findings, remediation, and retest outcomes.

Authors & Contacts



Mario Onorato
IBM Industry Diamond



Orazio Lascala
Director



Silvia Peschiera
Strategy and Transformation
Service Line Leader, Italy



Silvia Procacci
Enterprise Strategy Practice
Leader, Italy



Giulia Ugolini
Managing Consultant, Italy



Riccardo Cinelli
Managing Consultant, Italy



Matteo Muscolo
Strategy Consultant

IBM