

WHITEPAPER

When Agents Run the Bank

The End of the Second Line as You Know It

September 2026

By Matteo Coppola and Anne Kleppe

When Agents Run the Bank

The End of the Second Line as You Know It

Banks are deploying agentic AI systems that can interpret signals, invoke tools, sequence actions, and produce operational outcomes with far less human intervention than traditional automation or predictive models. This has significant implications for risk oversight because most banks still assume that review, challenge, and escalation can occur on human time, an assumption that becomes strained or even invalid in higher-autonomy workflows.

The strategic challenge for CEOs and Chief Risk Officers (CROs) is to ensure that the bank's governance model can keep pace as decision production moves from human-managed workflows to machine-orchestrated ones. Going forward the second line will need to shift from periodic reviews of human decisions to continuous governance of machine-orchestrated risk systems on a portfolio level.

This is a major structural redesign of the risk function that is a concern for CEOs as they consider AI implementations. In BCG's 2026 Global AI Risk Management Survey of about 30 leading international banks, 80% reported having an AI policy in place, yet one in five still reported issues or incidents of AI systems in production generating biased or incorrect outputs.

THE INFLECTION POINT

Reimagining Risk in the Age of Agents

Consider a tier-one bank using an agent for first-pass commercial loan monitoring across its mid-market portfolio. Within three months, it surveils tens of thousands of exposures and processes hundreds of reviews a day: analyzing cash flows, refreshing probability-of-default estimates, and flagging covenant concerns. Work that once required dozens of analysts now runs at agent speed and scale.

But imagine the agent slightly and systematically underweights one early-warning indicator in one sector. Each individual judgment is defensible, and each exception is within tolerance, but the bias repeats across thousands of exposures until it becomes a material risk position.

This is the governance challenge. Existing controls may identify design risks before deployment, but they cannot keep up once agents operate across high-volume workflows. Risk emerges from continuous execution, compounds through repetition, and remains hidden because each action appears reasonable until the aggregate pattern becomes visible.

Another approval layer will not help here. Governance has to be embedded in the system itself: safe-by-design agents, residual risk managed through control-by-code, continuous testing and validation, real-time monitoring, and defined human judgment gates.

PART ONE

Three Horizons, Three Different Governance Challenges

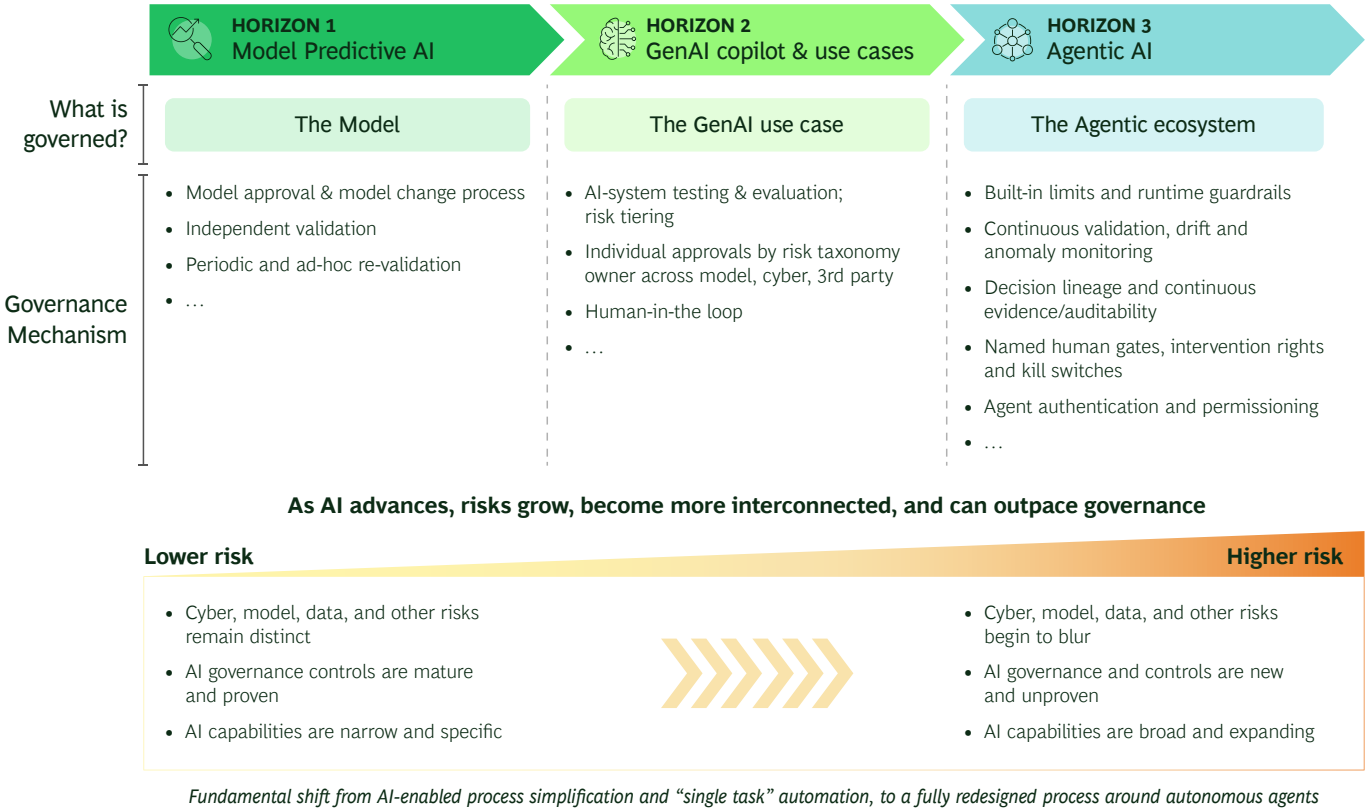
Banks are moving through three development horizons: from process automation and advanced analytics (Horizon 1) through GenAI copilots and use-case deployment (Horizon 2) toward agentic AI ecosystems in which autonomous agents execute end-to-end workflows, orchestrate across domains, and produce outcomes at a scale and speed no human review committee can match (Horizon 3). Even banks that decide not to deploy agents will face agentic risks because the market and their clients will be deploying them. (See exhibit 1).

Horizon 1 started a long time ago. It is the most familiar to banks, involving machine learning to make credit decisions or review transactions alerts in financial crime. Banks have model risk frameworks, validation routines, and supervisory expectations to govern it. The gap between what they have deployed and what they can govern is manageable.

Horizon 2 raises the stakes and started in the industry after 2022. Tools such as GenAI copilots, automated regulatory monitoring, and AI-generated credit memos can introduce hallucination, bias, and lead people to make decisions based on incorrect advice. These systems are general purpose (they can consume any input), they are unreliable (if you ask a question twice you will get two different answers) and unpredictable (if you change how you ask a question you can get drastically different outputs).

EXHIBIT 1

The Three Horizons of AI Development in Banking



Fundamental shift from AI-enabled process simplification and "single task" automation, to a fully redesigned process around autonomous agents

Source: BCG analysis.

Risks associated with horizons 1 and 2 are material but they are usually still governable using an enhanced and more continuous version of today's risk frameworks. But this framework should not be applied uniformly. Governance, testing, monitoring, and human-oversight requirements must be tailored to the system archetype, reflecting its specific failure modes, degree of autonomy, user interaction, and potential impact. For example, a chatbot cannot be tested like a traditional AI model; in particular back-testing must be part of a broader evaluation framework for ensuring the safety and security of GenAI systems.

Horizon 3 started very recently, we see the very first applications at our clients today and it is categorically different. An autonomous agentic system makes the monitoring call, invokes data services, updates the risk position, and generates the documentation across thousands of decisions a day. Its "model" is a dynamic composition of models, prompts, tools, data, memory, permissions, workflow logic and often human-machine interaction.

While bank leaders are accelerating toward Horizon 3, many others are stretching their risk frameworks built for Horizon 1 and adapted for Horizon 2 across AI deployments they were never designed to govern. The quarterly validation cycle common in Horizon 1 is already a challenge in Horizon 2, and certainly cannot be effective in Horizon 3 where it's necessary to govern a system that is operating continuously and composing decisions dynamically.

The Failure Modes Banks are Not Yet Measuring

Besides risks such as model error, data quality, algorithmic bias, banks must guard at least against three further serious problems that arise when systems act across workflows, tools, permissions, and time.

Drift. An agent's behavior can drift from its original intent with no obvious model change. What started as "flag suspicious activity" might become, over thousands of cases, "flag only what is clearly suspicious". Each decision is plausible but over time the pattern can create material risks.

Authorization failure. Most permission architectures at banks were built for people and applications, not agents. Running on employee credentials, an agent may inherit access that had been granted to a supervised human, reach the right system for the wrong purpose, combine permitted actions into an impermissible workflow, and leave an audit trail that identifies the credential but not the agent, delegation, or rationale.

Adversarial manipulation. Agents can be exposed to poisoned context, prompt injections, or malicious instructions embedded in documents, messages, or data feeds. The risk is that an agent executes a wrong action through a chain of individually permitted steps.

These risks show that agents must be governed as active participants in the operating environment, not passive models producing outputs. Pre-deployment validation and output review are not enough. They need behavioral baselines, purpose- and conduct-bound authorization, delegation controls, continuous evidence, and intervention protocols.

The Systemic Dimension

The risk extends beyond any single institution. As banks build agents on a concentrated AI supply chain (specialized hardware, cloud infrastructure, foundation models, and vendor tooling) they may import common dependencies, assumptions, and failure modes. If trading, treasury, or credit agents across major institutions interpret the same market signals through similar models or tooling, they could read the market the same way, move the same way, and amplify stress precisely when diversity of judgment is most needed.

This concern is already on the supervisory agenda. Both the Financial Stability Board and Bank for International Settlements's Financial Stability Institute have flagged AI-related service-provider concentration and market correlations as potential financial stability risks.[1]

PART TWO

A New Architecture for Risk: Seven Layers

Mapping today's committees and approval workflows onto an agentic operating model does not produce a working risk function. The design must start from the machine-orchestrated system that produces the outcomes.

The reference architecture has seven layers, each with a distinct governance logic, defined ownership, and clear interaction with adjacent layers (See exhibit 2). Its purpose is to show how to express existing principles (independent challenge, accountability, validation, monitoring, and audit) for a world in which agents increasingly make decisions rather than humans alone.

LAYER 1 — THE BANK'S WORLD MODEL

This layer answers one question: what is true about the bank and its environment right now? In practice, the world model is a governed ontology on top of the data platform: a knowledge graph that encodes what the bank's data means and how it connects (for example, that this customer is this counterparty, this obligor, under this covenant, with this control status and increasingly so also knowing that this agent is acting on behalf of this customer, the regulator etc.). Layer 1 lets agents reason over relationships.

In a traditional bank, this model exists in policies, procedures, and tacit knowledge held by thousands of people. In an agentic bank, the model must be explicit, integrated, and machine readable. The first line is accountable for the model's integrity because agents act on the business's operational facts. The second line does not own the model or vouch for every fact; it sets integrity standards and uses the same model to challenge the control framework.

LAYER 2 — GOVERNED RISK SERVICES (CAPABILITIES)

Layer 2 is the set of Risk and Compliance capabilities: credit assessment, KYC classification, sanctions and PEP screening, transaction monitoring, stress testing, control testing. These are defined around the risk function they perform, not the human workflows that historically wrapped them.

1. The Financial Stability Implications of Artificial Intelligence" (FSB, November 2024) — PDF: https://www.fsb.org/uploads/P14112024.pdf; BIS/FSI source "Regulating AI in the Financial Sector: Recent Developments and Main Challenges" — FSI Insights on Policy Implementation No. 63, BIS Financial Stability Institute, authored by Crisanto, Ehrentraud et al. (December 2024) — full paper: https://www.bis.org/fsi/publ/insights63.pdf (summary version: https://www.bis.org/fsi/publ/insights63_summary.pdf).

EXHIBIT 2

The Seven-Layer AI-Native Risk Architecture

ARCHITECTURE LAYER	DESCRIPTION	1st LINE (BUSINESS)	2nd LINE (RISK & COMPLIANCE)
L7 **Evidence & Reporting** (Explainability, Traceability, Provenance)	**Continuous, provable record** of **how decisions** were made and authorized	Generates **lineage & evidence**	Sets **evidence standards**; **reporting**
L6 **Human Judgment** (Verification & Accountability)	Defined **gates** where **named humans intervene**	Involves **named humans deciding & owning**	Intervenes at **defined gates**, **portfolio level** at the edge
L5 **Validation & Monitoring**	Independent **testing pre-launch**, then live **drift and anomaly monitoring**	Operates **continuous monitoring** of single agents	Owns **oversight over monitoring**
L4 **Risk Orchestration Engine** (Execution)	**Runtime** that composes governed services within **approved control limits**	Operates **engine within limits**	Sets **standards**
L3 **Policy-as-Code**	**Appetite and obligations** enforced as **machine-readable decision rules**	Implements **rules in code**	Owns **appetite, regulatory obligations & standards**
L2 **Governed Risk Services** (Capabilities)	**Risk and compliance capabilities** exposed as **governed, services**	Owns **services & results**	Sets **control standards**
L1 **World Model**	Live, governed **model of what is true** across the bank	Owns **model integrity & facts**	Sets **integrity standards & challenges**

Source: BCG analysis.

In practice, Layer 2 is a catalogue of APIs and services (screen-sanctions, assess-credit-risk, monitor-transaction) that a human or agent can call to make a structured request and receive a structured answer with evidence. The governed service is the full control object: the code, model, or rules it invokes, plus its schema, validation evidence, permissions, version history, monitoring, and audit trail.

This relocates where control lives. Today, a sanctions check is a button inside a case-management application, and the real control is the process around it. As a governed service, the check exists once, on its own, with validated thresholds, permitted uses, and required evidence built into it. The end state may not be fully agent-driven: some processes may remain largely deterministic, but be designed, assembled, adapted, and maintained by agents. Equally, some processes may become genuinely dynamic, with the orchestration engine composing the steps case by case. Control of each component becomes self-contained (Layer 2), while control of how components combine moves into machine readable rules (Layer 3) enforced at runtime (Layer 4). Every governed service has a named owner in the first line, the second line sets the standards a service must meet before it can be called.

Every governed service has a named owner in the first line, the second line sets the standards a service must meet before it can be called.

LAYER 3 — POLICY AS CODE

Layer 3 defines what the bank's own agents or agents acting on behalf of external parties must, may, and must never do. Risk appetite, regulatory obligations, and control standards become machine-readable decision rules, such as exposure thresholds that trigger human review, quality gates that outputs must clear, purpose and conduct limits on how a service can be used, and circuit breakers when error or drift breaches tolerance.

This is a policy engine for the obligations and appetite limits that must be enforced at decision time. Governance must be precise. The second line owns the risk appetite, regulatory obligations, materiality, and control standards. The first line including technology implements them under controlled change. In a document regime, a policy error propagates through human interpretation. In code, a miscoded exposure threshold can distort every credit decision before review detects the pattern.

LAYER 4 — RISK ORCHESTRATION ENGINE

This is where the first three layers meet. The orchestration engine delivers outcomes (credit decisions, KYC completions, suspicious-transaction investigations) by composing governed services dynamically, within Policy-as-Code controls. Coverage must be universal: all AI systems and agents, including business-owned ones, run through the engine and draw on the same governed services. More broadly, every individual bank's processes will be governed by this orchestration engine.

The engine should use deterministic logic wherever a case fits a known schema, reserving flexible reasoning for exceptions when the system cannot proceed safely. The second line sets the standards and determines the escalation framework while the first line, including technology, operates the engine within them.

LAYER 5 — VALIDATION & MONITORING

Layer 5 tests, independently, whether the system built in Layers 1 to 4 still behaves as intended. Model risk management is the closest analogue, but agentic oversight is broader. Since a single agent can create, model, and conduct operational, cyber, and compliance exposure in the same workflow, the company needs to monitor continuously across taxonomies.

Validation works in two modes. Before production, it establishes the agent's behavioral baseline, which includes expected input data, reasoning paths, boundary behavior, fairness profile, resilience under stress, and exposure to adversarial prompts or data leakage etc. Once the agent starts operating, the same layer tracks that baseline for drift, degradation, anomalies, and emerging conduct or security risk.

The second line owns the independent standards, challenge, and escalation thresholds.

LAYER 6 — HUMAN JUDGMENT (VERIFICATION & ACCOUNTABILITY)

Human judgment enters the system at defined points rather than sitting above it. It enters when Layer 4 determines the necessity for escalation (for example, when consequences require more ethical, reputational, or risk-appetite judgment than the bank feels safe transferring to the machine), and when Layer 5 detects drift, degradation or anomaly. A credit officer reviewing a $200 million exposure should not recalculate risk; the officer should decide whether the context of the edge-case changes the risk appetite—and accept accountability for that judgment.

Some serious AI control failures occur precisely at this human-in-the loop interface. People can be too trusting, reluctant to challenge, or inclined to rubber-stamp outputs if they cannot meaningfully interrogate them. This layer works only if verification is engineered into the workflow so that evidence, rationale, uncertainty, policy basis, source lineage, and escalation options are visible so that a reviewer is truly enabled to challenge the recommendation without being overburdened.

LAYER 7 — EVIDENCE & REPORTING

Layer 7 makes the work of the other six visible to management, regulators, and audit. It converts agentic activity into management reporting, exception workflows, audit interfaces, and supervisory evidence. Its purpose is to report outcomes and, critically, to prove how material decisions were produced and authorized.

For any material agentic decision, the bank should have the ability to show which agent acted, under whose delegation, using which credential, on which data, through which governed services, Policy-as-Code controls, and human judgment gates, and with what result.

Regulators will ask two questions. Can the bank explain an individual decision, and is the control environment itself governed, tested, monitored, and change controlled? Supervisory technology (SupTech) is moving supervision toward more automated reporting, richer supervisory analytics and, in some cases, real-time monitoring. Banks must move from reconstructing evidence after the fact to producing it continuously.

The seven-layer architecture should be dynamic by design, able to incorporate a change in regulation, risk appetite, or methodology end to end.

PART THREE

The Three Lines Redesigned

In the new risk management architecture, the first line's role remains similar. It owns the agents it deploys and remains accountable for their outcomes. But the second line transforms from reviewing human decisions to governing the architecture within which agents make decisions. It no longer approves individual AI systems or applications. It sets the standards for Policy-as-Code controls, orchestration limits, validation, monitoring, evidence, and world-model integrity without owning every fact or operating the engine. It has the authority to challenge the system directly, not to reperform individual decisions.

The third line may still use sampling to test controls, but it does not reperform agentic decisions as its primary assurance model. It tests whether the control architecture is well designed, operating as intended, and producing evidence that can be independently challenged. Did the orchestration engine follow approved composition rules? Is Policy-as-Code consistent with approved risk appetite and regulation? Can the bank demonstrate which agent acted, under whose delegation, for what purpose, within its conduct limits, and with what result?

People Are at the Edge

The new architecture changes what people do. In a traditional bank, much of the second line's work is spent moving context across systems, documents, committees: aggregating, escalating, aligning, documenting. The hierarchy remains essential for accountability, challenge, and escalation, but less of its value comes from moving information once the architecture carries more of that context itself.

The edge defines where risks emerge that Policy-as-Code has not encoded, where there is an unclear line of authority, when an agent is acting outside its purpose, or when a regulator, client, or Board needs a credible human answer The change shows up as three shifts in how existing capabilities are deployed.

The second line expertise is deployed no longer on static algorithms and focused on documentation and periodic reports but on hundreds of live components that must be continuously calibrated and monitored. This calls for risk differentiated oversight where most of the AI systems are approved and monitored at portfolio level rather than case by case. For example, the second line will have specialists who set and own the drift and anomaly thresholds for the consumer-credit book or conduct limits that the first line implements and runs.

Accountability needs to be reinforced for cross-cutting outcomes. Consider a prompt-injection attack that manipulates an AI-enabled customer risk assessment. Cyber owns the technical incident, Financial Crime owns the impact on customer due diligence, and Compliance owns the regulatory implications. The gap is that no single function may own the end-to-end response: identifying all affected decisions, coordinating remediation, deciding whether onboarding can continue, and confirming that the residual risk is acceptable.

Judgment shifts up toward what remains "human". As routine analysis and routing move to the architecture, senior expertise concentrates on technical judgment, challenge on edge cases, coaching, and developing the people who own the live risk components.

This shift will happen gradually. It will evolve from AI governance bolted onto existing risk silos, to a shared control backbone that lets all risk types draw on common data, policy, and validation standards, to a risk function organized around architecture and outcomes. Risk types will persist as expertise pools while the architecture handles coordination and people focus on depth of expertise, integration, and judgment.

CONCLUSION

The Leadership Choice

Agentic AI is coming. The question for risk leaders is whether the control infrastructure is in place before the volume arrives. Building it after is considerably more expensive.

About the Authors



Matteo Coppola
Global Leader for the Risk and Compliance Practice and a Managing Director and Senior Partner
BCG's Milan office
Coppola.Matteo@bcg.com



Anne Kleppe
Global Leader for Responsible AI and a Managing Director and Partner
BCG's Berlin office
Kleppe.Anne@bcg.com



About Boston Consulting Group

Boston Consulting Group bridges the gap between ambition and outcomes for the world's leading companies and organizations. We are built for this era of unprecedented change—bringing strategic clarity rooted in over 60 years of deep domain knowledge, combined with applied AI shaped by our practitioners. BCG works shoulder-to-shoulder with CEOs across industries and geographies to deliver transformative impact at scale: stronger returns, transferred capabilities, and change that sticks. For more information, visit bcg.com.

For information or permission to reprint, please contact BCG at permissions@bcg.com. To find the latest BCG content and register to receive e-alerts on this topic or others, please visit bcg.com. Follow Boston Consulting Group on LinkedIn, Instagram, Facebook, and X.

© Boston Consulting Group 2026. All rights reserved. 9/26


BCG