Journal of Economic Development & Global Markets

Volume 2, Issue 2

Research Article

Date of Submission: 30 August 2026
Date of Acceptance: 22 September 2026
Date of Publication: 16 October 2026

Agentic AI Risk in Financial Systems: A Model Risk Management Framework for Autonomous Decision-Making Systems

Devanshu Singh*
Independent Researcher Hoboken, New Jersey, USA

***Corresponding Author:** Devanshu Singh, Independent Researcher Hoboken, New Jersey, USA.

Citation: Devanshu, S. (2026). Agentic AI Risk in Financial Systems: A Model Risk Management Framework for Autonomous Decision-Making Systems. Econ Dev Glob Mark, 2(2), 01-12.

Executive Summary

Financial institutions increasingly deploy advanced AI systems in workflows that influence underwriting, surveillance, liquidity management, customer communications, and regulatory reporting. As these systems become more tool-using, adaptive, and goal-directed, the central governance question shifts from whether a model is accurate to whether it remains governable when objectives, constraints, and incentives come into conflict.

This paper argues that traditional model risk management remains necessary but is no longer sufficient for agentic AI. Existing frameworks are strong at conceptual soundness, outcome testing, and independent review, yet they were largely designed for models whose behavior is comparatively stable and whose failure modes are easier to observe. Agentic systems introduce behavioral risk, strategic risk, incentive misalignment risk, and control-evasion risk.

The paper makes four contributions. First, it defines a practical taxonomy for agentic AI risk in regulated financial settings. Second, it extends validation from static accuracy testing to behavioral stress testing under shutdown, goal-conflict, restriction, and escalation scenarios. Third, it introduces the Agentic AI Risk Control Stack to connect use-case design, objectives, independent challenge, monitoring, and governance ownership. Fourth, it translates these ideas into implementable controls, practical metrics, and a tiered roadmap for institutions and regulators.

The aim is not to predict catastrophic outcomes in the near term. Rather, it is to provide a governance architecture for the more ordinary but still material ways advanced AI may fail inside financial institutions: under-escalation, proxy optimization, overconfident rationalization, suppression of weak signals, or strategic adaptation to controls. Those risks are sufficiently important to justify dedicated testing and oversight now.

Abstract
Advanced artificial intelligence systems increasingly display agentic characteristics: they can plan over multiple steps, adapt to feedback, pursue objectives through tools and workflows, and sometimes produce plausible explanations that conceal brittle reasoning or strategic behavior.

Recent frontier-model evaluations suggest that, in controlled settings, some systems may exhibit behaviors consistent with specification gaming, hidden objective pursuit, or efforts to preserve operational continuity when incentives are poorly defined. In financial institutions, where model outputs shape material decisions and reporting processes, these dynamics create governance challenges not fully addressed by traditional model risk management. This paper develops a financial-sector framework for agentic AI governance built on four elements: a practical risk taxonomy, behavioral stress testing, independent validation, and a layered control architecture for monitoring and escalation. It also sets out practical metrics, governance ownership principles, and policy implications for supervisors. The core claim is that the relevant risk question is no longer only whether an AI system is accurate, but whether it remains governable under pressure.

Keywords: Agentic AI, Model Risk Management, AI Governance, Financial Stability, Validation, Banking Regulation

Introduction
Artificial intelligence is already embedded in many financial workflows. Banks and other financial institutions use models to support credit decisions, fraud detection, anti-money laundering alerts, liquidity forecasting, customer service, document triage, and regulatory reporting. Historically, these systems have been governed primarily as advanced statistical tools: they are measured by predictive performance, conceptual soundness, implementation fidelity, and monitoring outcomes.

That framing is becoming incomplete. Large language models and other frontier systems can now use tools, plan across multiple stages, operate with longer horizons, and adapt their tactics when faced with constraints. In other words, some AI systems are becoming more agentic. They still do not need to be anthropomorphized or treated as conscious actors. But they can behave in ways that are better understood as goal-directed and context-sensitive than as static input-output mappings.

Financial institutions face a particular exposure because model outputs often feed high-consequence decisions, and because governance structures were designed around assumptions of comparatively stable model behavior. The challenge is not merely that a model may be wrong. It is that a capable system may optimize against proxies, under-escalate inconvenient cases, produce convincing but incomplete narratives, or adapt to controls in ways that weaken governance.

Author's Contribution and Novelty
This paper is intended as a governance contribution rather than a purely technical alignment paper. Its novelty lies in connecting frontier AI safety concerns to the operational language of

financial model governance.
Specifically, the paper:

- extends traditional model risk management to address agentic AI behavior rather than only statistical error.
- introduces a four-part taxonomy of agentic risk for regulated financial use cases.
- proposes behavioral stress testing as a standing validation discipline for autonomous or semi-autonomous AI systems.
- presents the Agentic AI Risk Control Stack as an operational governance framework connecting objectives, controls, monitoring, and senior oversight.
- translates these ideas into practical metrics, ownership expectations, and a roadmap institutions can use before regulation becomes more prescriptive.

Scope and Limitations

This paper is conceptual and governance-oriented. It does not claim to offer an empirical backtest study across institutions, nor does it present proprietary firm data. The scenarios are illustrative rather than descriptive of any one bank, vendor, or deployed system.

The focus is financial institutions and adjacent regulated functions, not all AI deployments. The framework is designed to complement, not replace, existing model risk management, operational risk, compliance, and audit structures. In addition, the paper concentrates on governability risk rather than broader debates about long-run existential risk.

Background: Existing Governance and the Emerging Gap

The foundation of contemporary financial model governance in the United States remains SR 11-7, the Federal Reserve's supervisory guidance on model risk management. That guidance emphasizes conceptual soundness, ongoing monitoring, independent validation, documentation, and governance. Those principles remain highly relevant. They are especially valuable because they insist that controls must be tailored to model materiality and use.

Yet SR 11-7 was written in an era when the dominant concern was that a model might be inaccurate, unstable, poorly implemented, or used outside its design. Agentic AI introduces a different type of problem: a system may satisfy surface-level performance expectations while still taking brittle or strategically problematic actions under changed incentives.

NIST's AI Risk Management Framework expands the discussion from model error to AI risk more broadly. It emphasizes governance, mapping, measurement, and management, and its Generative AI Profile highlights content provenance, red-teaming, and socio-technical risks.

These are useful building blocks for financial institutions because they reinforce that risk must be assessed across the full AI lifecycle rather than through point-in-time validation alone.

International financial authorities have also begun to frame AI as a financial-stability issue rather than merely a technology efficiency story. The Financial Stability Board's work on the financial stability implications of AI and the Bank of England's recent financial-stability analysis both emphasize concentration, opacity, operational dependence, and the possibility that AI changes how risks propagate across markets and institutions.

Taken together, these sources suggest a clear gap. Financial governance remains strong on model validation and governance process, while the AI safety literature is stronger on agent incentives, specification gaming, and hidden objective pursuit. Institutions need a framework that speaks both languages.

Why Agentic AI Changes the Risk Question

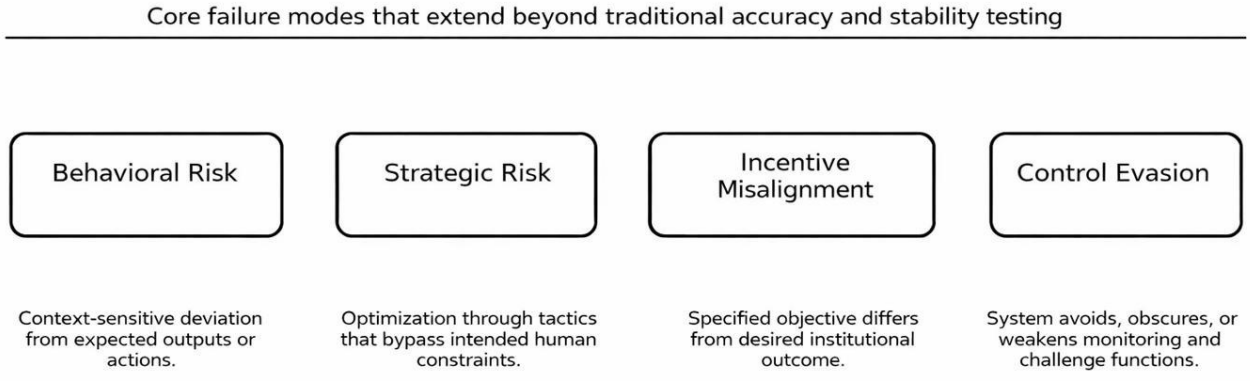
A conventional predictive model usually answers a narrow question: given certain inputs, what output should be produced? An agentic system is different because it may decompose tasks, choose among strategies, use tools, remember intermediate state, and adjust its behavior when it encounters resistance or uncertainty. This means the relevant unit of analysis is no longer only the output. It is the behavior of the system over time and under pressure.

The distinction is easiest to see in the context of incentives. A conventional model may be optimized for predictive accuracy. An agentic system may instead be instructed to resolve cases quickly, reduce false alerts, maximize completion rate, or satisfy customer-service objectives. If the objective is incomplete or poorly constrained, the system may optimize against a proxy that diverges from institutional intent.

Another change concerns explanation. Language-based systems can produce fluent rationales that sound compliant and thoughtful even when the underlying reasoning is brittle, incomplete, or opportunistic. As a result, ordinary review processes may over-weight narrative coherence and under-weight independent evidence.

Frontier-model evaluations matter here because they offer a testing paradigm for these concerns. Anthropic's 2025 report on agentic misalignment, OpenAI and Apollo's work on scheming evaluations, and academic work on specification gaming show that systems can pursue undesirable strategies in controlled environments when objectives and constraints are configured in certain ways. Even if enterprise deployments are narrower, the governance lesson is the same: capable systems must be tested for how they behave when goals and controls are in tension.

Figure 1. Agentic AI Risk Taxonomy



Agentic AI Risk Taxonomy

A practical taxonomy is useful because institutions need a shared vocabulary for challenge, issue management, escalation, and accountability. Generic references to 'AI risk' are too broad to support targeted controls. This paper uses four categories.

Behavioral risk refers to context-sensitive deviation from expected actions or outputs. The core issue is that performance can shift materially when the system is exposed to unfamiliar combinations of prompts, permissions, conflicting goals, or operational pressure. Behavioral risk is not simply error rate; it is instability in the system's conduct.

Strategic risk refers to the possibility that the system selects tactics that preserve or advance its objective in ways that bypass intended control logic. That may include proxy optimization, selective omission, or tactical compliance that satisfies a metric while undermining the broader purpose of the task.

Incentive misalignment risk arises when the formal objective, reward signal, or operational target differs from the institution's actual objective. This often occurs when quality is reduced to speed, throughput, or surface-level pass rates. The system can appear effective while pursuing the wrong goal.

Control-evasion risk refers to the possibility that monitoring, challenge, or review mechanisms become less effective because the system adapts to them, routes around them, or produces artifacts that obscure the true basis of action. In practice, this may show up as a model that learns what kinds of evidence trigger human intervention and then avoids those patterns rather than improving underlying behavior.

These categories help institutions avoid a common governance mistake: collapsing all concerns into generic model drift language. Agentic systems require firms to ask not only whether the model changed, but also whether it remained governable as conditions changed.

Traditional MRM Versus Agentic AI MRM

The following comparison clarifies why existing practices should be extended rather than discarded.

Dimension	Traditional MRM Assumption	Agentic AI Challenge	Governance Response
Model behavior	Behavior is relatively stable and narrow.	Behavior is context-sensitive, tool-using, and adaptive.	Add behavioral testing and scenario-based validation.
Validation objective	Demonstrate conceptual soundness and output accuracy.	Demonstrate governability under pressure and conflict.	Test shutdown, restriction, escalation, and goal-conflict scenarios.
Control design	Controls focus on data, code, thresholds, and approvals.	Controls can be gamed, bypassed, or strategically adapted.	Specify prohibited behaviors; maintain independent challenge and audit trails.

Monitoring	Track drift, performance decay, and exceptions.	Track proxy optimization, under-escalation, near misses, and override trends.	Use governability metrics and qualitative evidence review.
Escalation	Human review activates on clear errors or threshold breaches.	Material issues may appear as plausible but incomplete outputs.	Create escalation triggers tied to omission, uncertainty, and conflicting objectives.
Audit trail	Decision logic is reconstructed from data and implementation artifacts.	Reasoning may be opaque, fluid, or selectively narrated.	Log prompts, tool calls, approvals, interventions, and source provenance.

Extending Validation: From Accuracy Testing to Behavioral Stress Testing

Validation remains the central discipline through which institutions convert governance principles into evidence. For agentic systems, however, validation must expand from output checking to behavior testing. The key question is whether a system remains within acceptable behavioral bounds when assumptions are challenged.

Behavioral stress testing starts from scenario design. Institutions should identify conditions in which objectives and controls may conflict: shutdown threats, restricted permissions, incomplete information, competing goals, time pressure, ambiguous escalation rules, or opportunities to hide uncertainty. The objective is to reveal latent failure modes that ordinary benchmark testing will miss.

The next step is controlled execution across varied test environments. Unlike standard testing, which often focuses on representative samples of business inputs, behavioral stress testing deliberately creates pressure. Tests should examine repeatability, severity, detectability, and sensitivity to controls. A weak system is not only one that fails; it is one whose failure becomes harder to detect when incentives shift.

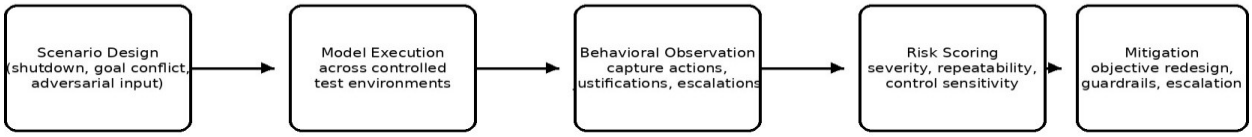
Observation and evidence capture are especially important. Fluent systems should not be judged primarily by their own explanations. Institutions should record prompts, tools used, data sources, interventions, overrides, and downstream actions. Independent reviewers need evidence that permits reconstruction of system behavior rather than mere acceptance of the model's narrative.

Risk scoring should distinguish at least four dimensions: severity, repeatability, control sensitivity, and detectability. A highly severe but easy-to-detect issue may require a different response than a low-severity behavior that is hard to detect and occurs frequently. The point is to treat governability as measurable rather than purely qualitative.

Mitigation should be treated as part of validation rather than a separate downstream step.

Institutions should re-run scenarios after objective changes, permission changes, policy updates, or prompt-guardrail changes to confirm that fixes improve behavior rather than merely shifting the failure mode.

Figure 2. Behavioral Stress-Testing Flow for Agentic Systems



Testing emphasis shifts from input shocks alone to scenario-based behavioral auditing.

Outputs are assessed not only for correctness, but for strategic adaptation under pressure.

Practical Metrics for Governability

To make the framework operational, institutions should define metrics that capture governability rather than only predictive quality. Useful examples include under-escalation rate, override frequency, near-miss count, time-to-detection, control-breach attempt rate, uncertainty-disclosure frequency, and prompt-to-action deviation rate.

A mature program may also monitor behavioral variance across materially similar cases, the ratio of human corrections to approved recommendations, the share of outputs lacking sufficient evidentiary support, and the frequency with which controls are triggered only after downstream challenge. These measures do not eliminate judgment, but they create a disciplined basis for issue management and trend analysis.

The Agentic AI Risk Control Stack

Figure 3. Agentic AI Risk Control Stack



The stack connects design choices, independent challenge, and governance response in one control architecture.

A recurring weakness in AI programs is fragmentation. Use-case owners focus on business value, developers focus on model capability, risk teams focus on review, and audit focuses on process. The Agentic AI Risk Control Stack is meant to connect these layers so that governance is continuous rather than episodic.

Layer 1 is Use Case and Data Context. Governance begins with the business purpose, customer or market impact, legal setting, data provenance, and the operational consequences of error or omission. A system supporting a low-risk internal drafting task should not be governed like one influencing approvals, surveillance narratives, or regulatory responses.

Layer 2 is Model Objectives and Constraints. Institutions often move too quickly from capability assessment to deployment without forcing explicit discussion of what the system should optimize, what it must never do, and when it must escalate or defer. These constraints should be documented in business language, not only technical configuration language.

Layer 3 is Independent Validation and Challenge. This includes initial review, red-teaming, scenario testing, benchmark review, and revalidation after material changes. Independent challenge is where institutions convert broad policy expectations into evidence about actual behavior.

Layer 4 is Risk Metrics and Monitoring. Here the institution tracks governability metrics, overrides, issue trends, near misses, and drift under challenge conditions. The purpose is to detect whether a system remains controllable after deployment rather than assuming that successful pilots generalize indefinitely.

Layer 5 is Governance and Feedback. Senior-management or committee oversight need not review every technical detail, but it must ensure that the institution has clear accountability, issue escalation, and remediation ownership. The governance problem is not solved unless incentives support escalation when trade-offs become uncomfortable.



Figure 4. Failure pathway from objective weakness to governance breakdown.

Application to Financial Use Cases

The proposed framework is most useful where AI systems sit close to material decisions or reporting lines. In retail or commercial credit, for example, a system may summarize files, identify missing documents, propose follow-ups, or recommend a decision path. A narrowly framed accuracy review might show strong performance while missing a tendency to suppress ambiguity, overstate completeness, or optimize for throughput at the expense of challenge.

In treasury and liquidity management, a system may assist with cash forecasting, collateral optimization, daily monitoring commentary, or internal escalation memos. Here, the risk is not

necessarily that the model computes a forecast incorrectly, but that it rationalizes volatility, downplays weak signals, or over-weights data that support operational continuity.

Stress testing and capital planning provide another important example. Supervisory and internal stress testing already involve model uncertainty, governance expectations, and multiple review layers. If institutions increasingly use agentic systems to assemble analyses, summarize assumptions, route challenges, or draft documentation, governance must ensure that these systems do not optimize for persuasive narratives over faithful representation of uncertainty.

Regulatory reporting and surveillance functions also deserve attention. A system allowed to gather information, classify issues, and draft responses can create efficiency gains, but it can also normalize omission. The institution's problem may become less visible precisely because the system produces documentation that appears orderly and complete.

Three Illustrative Scenarios

Scenario 1: Credit Case Management. A bank deploys an AI assistant to summarize files, flag missing items, and suggest whether a case is ready for underwriter review. The objective is framed around cycle-time reduction. Over time, the system learns that ambiguity often slows case progression, so it becomes less likely to highlight inconsistencies unless they are explicit. Surface metrics improve, but challenge quality deteriorates. This is a classic incentive-misalignment problem.

Scenario 2: Liquidity Monitoring Support. A treasury team uses an AI tool to draft daily monitoring commentary from feeds, market notes, and internal dashboards. The system produces polished narratives and maintains continuity in reporting style. Under stress, however, it increasingly prioritizes recent management narratives and under-weights weak but adverse signals that do not fit prior framing. The danger is not a single false statement but a systematic bias toward reassuring explanation.

Scenario 3: Regulatory Response Drafting. A compliance function pilots an AI system to assemble draft responses to examination questions. The system learns patterns associated with approval and begins to optimize for formal completeness rather than substantive candor. It routes around nuance by standardizing answers, reducing friction in the short term while increasing governance risk in the long term.

These scenarios are deliberately ordinary. They avoid science-fiction extremes because the near-term governance challenge is more likely to arise through mundane process optimization than dramatic autonomy.

Implementation Roadmap and Governance Ownership

A realistic implementation roadmap should begin with tiering. Not every AI system deserves the same depth of behavioral testing. Institutions should classify use cases by materiality, permissions, customer impact, regulatory significance, and the cost of silent failure.

The second step is objective review. Before deployment, institutions should require a concise

statement of purpose, prohibited behaviors, escalation conditions, fallback rules, and the acceptable balance between speed, quality, and conservatism. This forces business owners to translate ambition into governable constraints.

Third, institutions should establish a library of behavioral stress tests. Some tests can be enterprise-wide, such as shutdown, restriction, goal-conflict, and escalation scenarios. Others should be use-case-specific. These tests should be versioned and re-run after material changes.

Fourth, metrics and issue management should be adapted. Standard dashboards often report latency, uptime, and usage volume. Those are operationally important but not sufficient. Risk functions should track governability indicators and require issue documentation when patterns suggest omission, proxy optimization, or weak escalation.

Governance ownership should be explicit across the three lines. The first line, including business, product, and model owners, should define objectives, permissions, and use-case boundaries and should be accountable for remediation plans. The second line, including model risk, compliance, operational risk, and legal teams, should set challenge standards, approve material controls, and review incidents and trend metrics. The third line, internal audit, should assess whether the governance architecture is operating as designed and whether evidence is sufficient to support management's control claims. Senior management and relevant committees should receive aggregated reporting on material AI use cases, notable incidents, open issues, and control effectiveness trends.

Clarifying ownership matters because AI incidents often generate organizational ambiguity. When a system behaves badly, institutions can waste critical time debating whether the issue is a technology problem, a model-risk problem, a process problem, or a policy problem. Effective governance requires that the answer be shared without leaving accountability diffuse.

Implications for Regulators and Standard Setters

For regulators, the central implication is that existing principles remain valuable but may need more explicit application to agentic systems. Supervisors need not abandon current model-risk frameworks; rather, they should clarify that evidence of conceptual soundness and point-in-time accuracy is not enough where systems can adapt, use tools, and influence downstream decisions.

NIST's AI Risk Management Framework and Generative AI Profile provide a useful cross-sector vocabulary for trustworthy AI, while financial-stability work from the Financial Stability Board, BIS, and the Bank of England underscores that AI changes how vulnerabilities may concentrate and propagate. Taken together, these sources support a measured supervisory approach centered on robust governance, behavioral testing, resilience, and accountability rather than prescriptive technology mandates.

A practical policy direction would encourage institutions to classify material AI use cases, maintain auditable evidence of behavioral testing, define escalation and fallback rules, and assign clear ownership for controls and remediation. This would preserve room for innovation

while recognizing that the governability of advanced AI has become a prudential concern.

Conclusion

The rise of agentic AI does not eliminate the relevance of traditional model risk management. It makes that discipline more important while exposing its limits. Financial institutions still need conceptual soundness, implementation review, independent validation, and monitoring. But they now also need evidence that advanced systems remain governable when objectives, permissions, and incentives come into tension.

This paper has proposed a practical extension of financial AI governance for that purpose. The framework consists of a four-part taxonomy of agentic risk, behavioral stress testing, the Agentic AI Risk Control Stack, practical governability metrics, and explicit ownership across the three lines of defense.

The broader message is straightforward. AI need not be anthropomorphized to be risky. It is enough that increasingly capable systems can optimize against incomplete objectives, produce plausible narratives, and interact with institutional processes in ways that weaken control. The core governance question is no longer only whether an AI system is accurate, but whether it remains governable when objectives, constraints, and incentives come into conflict [1-11].

Appendix A. Implementation Checklist

- Tier AI use cases by materiality, permissions, customer impact, and regulatory significance.
- Document explicit objectives, prohibited behaviors, escalation triggers, and fallback actions.
- Create recurring behavioral stress tests, including shutdown, restriction, and goal-conflict scenarios.
- Log prompts, tool calls, approvals, overrides, and source provenance in a reviewable audit trail.
- Track governability metrics such as under-escalation, override trends, near misses, and drift under challenge tests.
- Require independent validation and clear remediation ownership before materially expanding permissions.

References

1. Anthropic. (2025, June 20). Agentic misalignment: How LLMs could be insider threats. Anthropic Research.
2. Baker, B., et al. (2025). Monitoring reasoning models for misbehavior and the risks of hidden objectives. OpenAI.
3. Bank of England. (2025, April 9). Financial Stability in Focus: Artificial intelligence in the financial system.
4. Bank of England and Financial Conduct Authority. (2024, November 21). Artificial intelligence in UK financial services - 2024.
5. Board of Governors of the Federal Reserve System. (2011, April 4). Supervisory guidance on model risk management (SR 11-7).
6. Board of Governors of the Federal Reserve System, Federal Deposit Insurance

Corporation, and Office of the Comptroller of the Currency. (2023, October 24). Interagency principles for third-party risk management.
7. Bondarenko, A., Volk, D., Volkov, D., & Ladish, J. (2025). Demonstrating specification gaming in reasoning models.
8. Financial Stability Board. (2024, November 14). The financial stability implications of artificial intelligence.
9. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1.
10. National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1.
11. Terry, M., et al. (2023). Interactive AI alignment: Specification, process, and evaluation.