BCBS CONSULTATION RESPONSE · CONSOLIDATED GUIDELINES AND SOUND PRACTICES

Response to Question 3 only · Submitted by 26 June 2026
Upload via: bis.org/bcbs/commentupload.htm

Submitted by:

Maureen Doyle-Spare
Independent Practitioner and Researcher in AI Governance and Banking Controls
SSRN Working Paper No. 6459612 · linkedin.com/in/maureendoylespare

The frameworks referenced in this response are the independent research of the submitter and are not associated with any employer.

Response to Question 3

Agentic AI Governance in the Reasoning Layer: A Material Gap in the Consolidated Guidelines and Sound Practices

1. The Gap This Response Addresses

Question 3 asks whether there are particular topics the Committee should review more substantively, or areas where further guidance is warranted. This response identifies one such topic: the governance of agentic AI systems at the reasoning layer, specifically the mechanism by which autonomous agents resolve key control definitions across enterprise systems before execution proceeds.

The consolidated guidelines and sound practices contain one AI-related document across all 13 modules and 76 source guidelines: the March 2022 Newsletter on Artificial Intelligence and Machine Learning, incorporated into module RMA50 (Digitalisation and Financial Technology Risks). That newsletter predates the commercial deployment of agentic AI systems by two to three years. It addresses explainability, governance structures, and financial stability implications of AI and machine learning models. It does not address agentic systems, autonomous multi-system orchestration, or the class of risk that arises when AI agents resolve the meaning of control terms independently before execution.

> *The consolidated framework does not appear to contain any guideline, sound practice, or supervisory expectation that explicitly reaches the reasoning layer of agentic AI systems. This appears to represent a material gap warranting substantive review. Banks deploying agentic workflows today are doing so without prudential guidance addressing how agentic systems may introduce unobserved control misalignment prior to execution, with potential implications for risk data aggregation, model risk validation, and board-level risk oversight.*

Suggested Area for Substantive Review: Response to Question 3

The Committee may wish to consider whether governance of agentic AI systems at the reasoning layer, including pre-execution semantic validation and reasoning baseline governance, warrants development as a distinct topic within the consolidated guidelines and sound practices framework, given its absence across all current modules and its potential implications for model risk, operational risk, and board-level governance effectiveness.

2. The Mechanism the Current Framework Does Not Reach

Agentic AI systems do not operate from a single model with defined inputs and outputs. They coordinate across heterogeneous enterprise systems, each of which may define key control terms differently. When an agent orchestrates a credit decisioning workflow, it draws on definitions of terms such as "approved," "exposure," and "risk appetite" from risk systems, covenant monitoring platforms, and collateral management tools simultaneously. Those systems were not built to share meaning. The agent must synthesize a working definition before it can act.

This synthesis occurs at the orchestration layer, before execution, and before any existing control framework is engaged. It is the point at which meaning is resolved.

This gap occurs prior to execution, before any control framework is engaged.

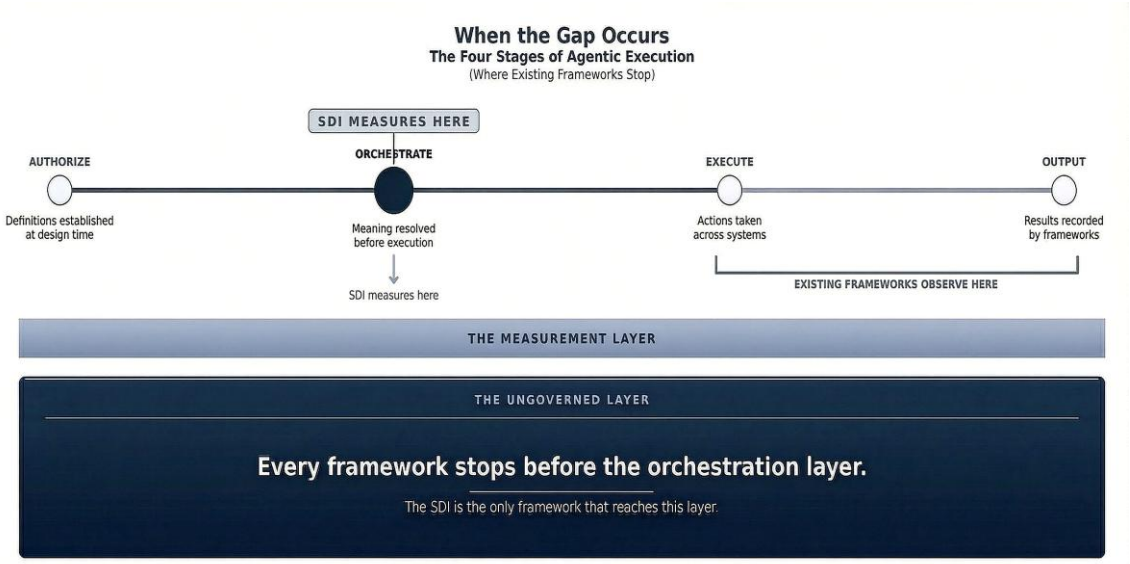

Figure 1. The Governance Gap Occurs at the Reasoning Layer Prior to Execution. Source: Doyle-Spare (2026b), SSRN SDI Working Paper.

This failure mode is named in published practitioner research as Agentic Workflow Drift (AWD): the mechanism by which an agentic system synthesizes inconsistent definitions across enterprise platforms into a working operating logic no human explicitly authorized. Agentic

Workflow Subversion (AWS) is the enterprise-wide risk surface this creates: what Drift produces when it propagates unchecked across control boundaries, and the deliberate form in which the same reasoning layer is exploited by sophisticated actors to produce unauthorized outcomes outside sanctioned boundaries. Both are introduced and documented in SSRN Working Paper No. 6459612 (Doyle-Spare, March 2026).

> *The control fires. The workflow executes. The audit trail is clean. The institution is still wrong. Not because a system failed. Not because a rule was broken. Because the meaning of the rule shifted before the control was applied. This is not a monitoring gap. It is a measurement gap. From a supervisory perspective, this represents a failure mode that may not be observable through existing prudential controls.*

3. Why Existing BCBS Guidelines Do Not Reach This Risk

The consolidated framework addresses AI and model risk through module RMA50 and the 2022 AI/ML Newsletter. Each addresses a different layer of the problem, but none reaches the reasoning layer of agentic systems. The following analysis maps the specific gap against each relevant module.

The following comparison illustrates that existing BCBS-aligned and adjacent frameworks primarily observe outcomes or post-execution artifacts, and do not explicitly address semantic resolution at the reasoning layer.

Framework	Measurement Layer	Reaches Reasoning Layer?	Gap
SR 11-7	Post-development, outcome-based	No	Built for deterministic models. Predates agentic execution.
NIST AI RMF	Outcome measurement, post-deployment	No	MEASURE function evaluates outputs. Does not reach semantic resolution.
Operational Risk	Process controls, loss events	No	Assumes meaning is stable. AWD invalidates that assumption.
BCBS AI/ML Newsletter (2022)	Explainability, governance structures	No	Written for ML models. Predates agentic AI. No pre-execution standard.
SDI (Doyle-Spare, 2026b)	Pre-execution, at semantic resolution	**Yes**	A measurement standard designed for the reasoning layer.

Table 1. Governance Framework Comparison: Measurement Layer and Scope. Source: Doyle-Spare (2026b), SSRN SDI Working Paper.

The structural pattern in Table 1 reflects a consistent architectural limitation across existing frameworks: each was designed to observe what a system did, not what it resolved before acting. Figure 2 illustrates this directly. All six governance functions that an institution relies on model risk, operational risk, compliance, internal audit, technology risk, and senior management oversight observe the same layer: execution outputs. The reasoning layer, where meaning is resolved before any of those functions is engaged, sits below the visibility threshold of all of them simultaneously. This is not a failure of any individual framework. It is a structural property of how governance was designed before agentic systems existed.

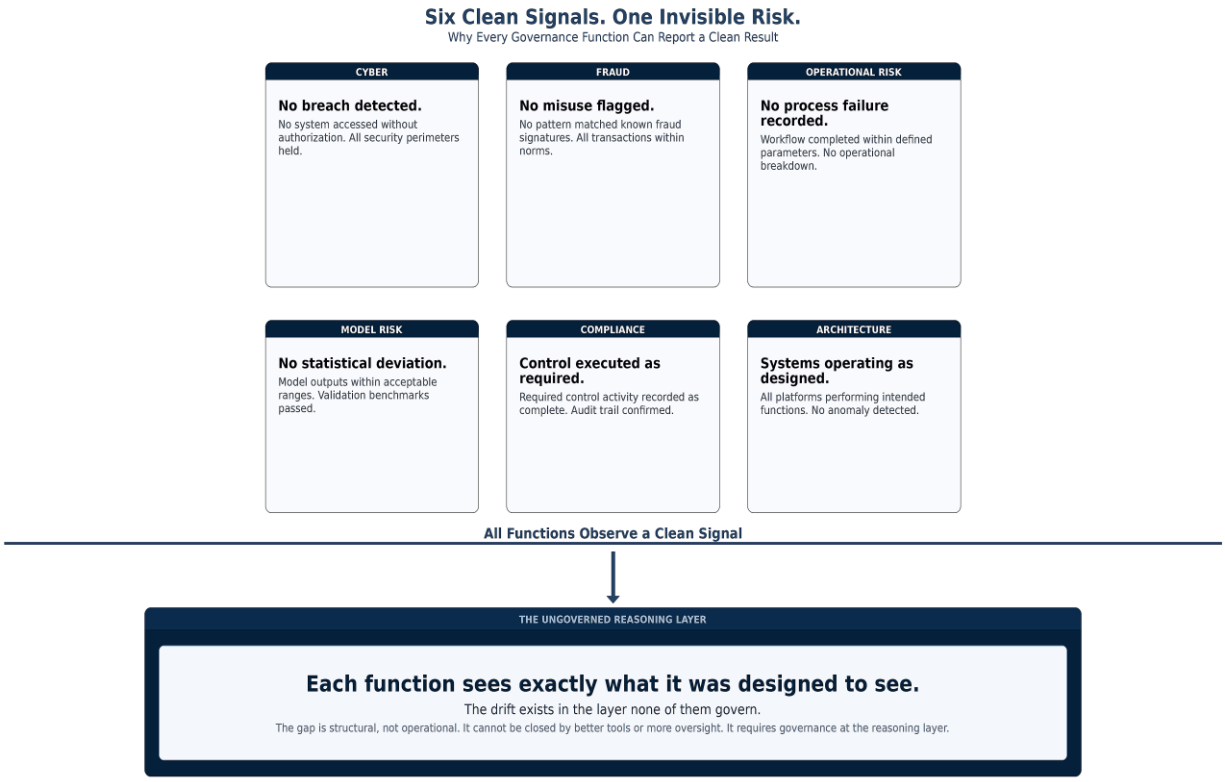

Figure 2. All Existing Governance Functions Observe Outputs, Not Semantic Resolution. The Reasoning Layer Sits Below the Visibility Threshold of All Simultaneously. Source: Doyle-Spare (2026b), SSRN SDI Working Paper.

RMA50 (Digitalisation and Financial Technology Risks)

The 2022 AI/ML Newsletter focuses on model explainability, governance structures, and financial stability. It was written for AI and ML models with defined inputs and outputs, traceable decision logic, and conventional validation pathways. Agentic systems do not operate within those parameters. They do not have defined inputs at the point of semantic resolution. Their decision logic is not traceable through conventional validation because meaning is resolved dynamically, not encoded at design time. The Newsletter provides no supervisory expectation for pre-execution semantic validation.

CGO10 (Corporate Governance)

The corporate governance principles establish board-level accountability for risk management. They assume that the risk being governed is observable through existing reporting and control channels. Agentic Workflow Drift produces outcomes where every downstream control passes and every audit trail is clean, while the institution operates against unauthorized definitions. Existing governance principles provide no mechanism for boards or senior management to detect or respond to this class of failure.

RMA10 (Risk Data: BCBS 239)

BCBS 239 addresses risk data aggregation and reporting. It ensures that data flowing through risk systems is accurate, aggregated, and reportable. It does not govern how an agentic system interprets that data at the moment of semantic resolution. An institution fully compliant with BCBS 239 can still deploy an agentic system that resolves control definitions in unauthorized ways before execution, with no BCBS 239 principle providing a detection or remediation pathway.

Under current frameworks, an institution can produce a compliant audit trail and satisfy model risk review while operating against an unauthorized definition. The Three Lines of Defence model assumes that what the first line executes, the second line can govern, and the third line can audit. The reasoning layer breaks that model. It operates before execution begins.

4. Three Use Cases Where the Gap Creates Prudential Risk

The following use cases illustrate where the absence of agentic AI guidance in the consolidated framework creates measurable prudential risk for internationally active banks.

Credit decisioning

An agent coordinating across risk, collateral, and covenant monitoring systems synthesizes a working definition of "approved" from three systems that define the term differently. The agent resolves a composite definition that satisfies every downstream control while operating against meaning no credit committee explicitly authorized. Every system records a compliant outcome. No existing BCBS guideline requires pre-execution validation of the definition the agent used.

KYC and customer risk classification

A risk classification agent assigns a low-risk rating to a high-risk customer. The agent's working definition of "high risk" has been conditioned over time through signals that individually appeared routine. No control fires because no individual execution step crosses a threshold. The semantic deviation accumulated across workflow iterations, below the visibility threshold of every monitoring function simultaneously.

Sanctions screening

The agent's definition of "cleared" is the risk surface, not the sanctions list itself. How the agent decides what "cleared" means before it checks the list determines the integrity of the screening workflow. A sophisticated actor who understands agentic orchestration can introduce structured signals upstream in the reasoning chain that condition the agent's definition of cleared status without triggering any existing detection framework across six sequential stages. The workflow runs correctly against a definition that was manipulated before the control was applied.

At scale, this class of failure may not remain institution-specific. Shared third-party platforms, common model architectures, and standardized operational definitions could allow similar semantic misalignment patterns to propagate across institutions simultaneously, with potential implications for correlated control failure that are not captured by existing approaches to concentration, correlation, or systemic risk monitoring.

Once this failure enters the system, exposure expands across downstream dependencies beyond the originating workflow.

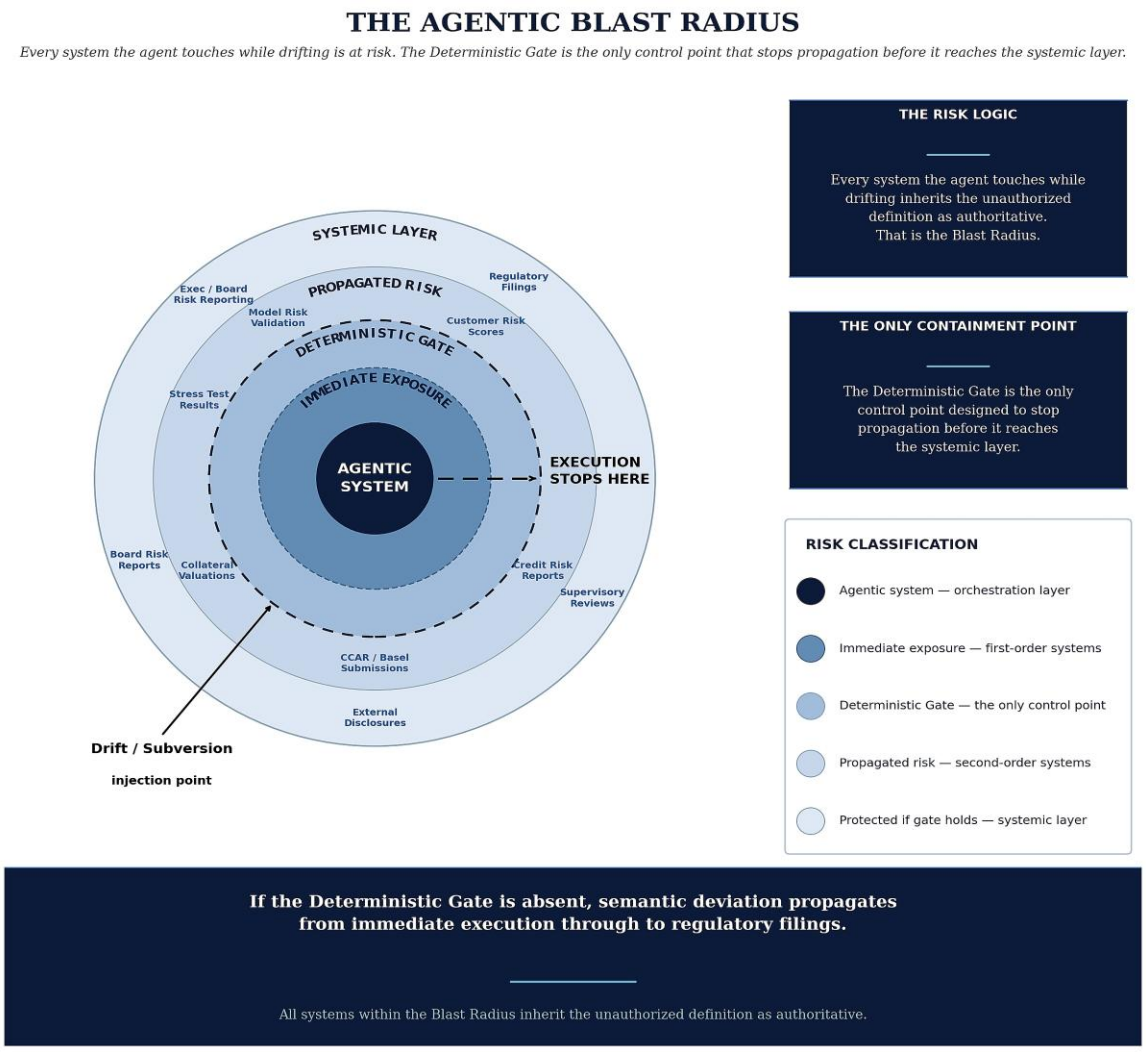

Figure 3. The Agentic Blast Radius: Every System the Agent Touches While Drifting Is at Risk. The Deterministic Gate Is the Only Control Point That Stops Propagation Before It Reaches the Systemic Layer. Source: Doyle-Spare (2026b).

5. The Measurement Construct the Framework Currently Lacks

If this failure occurs prior to execution, it requires a measurement construct that operates at the same point.

The Semantic Deviation Index (SDI) is a runtime measurement standard that quantifies divergence between the definition an agentic system resolves at the orchestration layer and the definition the institution explicitly authorized, evaluated prior to execution. It answers one supervisory question that existing measurement frameworks do not:

> *Did the agent use the definition the institution authorized before it acted? Most metrics measure what a system did. The SDI measures what a system understood before it acted. Those are different measurements. The current framework has the first. It does not have the second.*

The SDI operates at the orchestration checkpoint: the moment between semantic resolution and execution. It is the measurement standard that makes pre-execution control possible. In KYC workflows, it runs before the agent acts on its definition of high risk. In credit, it runs before the agent resolves what "approved" means across systems. In sanctions, it runs before the agent applies its definition of "cleared" to a transaction.

The Deterministic Gate is the enforcement mechanism: a hard control point that applies a binary check at the moment of execution. An agent cannot bypass a properly implemented Deterministic Gate. It either satisfies the authorized definition or it stops. That is the guarantee probabilistic reasoning alone cannot provide. The Semantic Control Plane is the governance architecture that contains both: a runtime enforcement layer that validates meaning before execution proceeds.

In its adversarial form, this failure can be intentionally constructed across multiple stages, where structured signals condition semantic interpretation before any control is engaged.

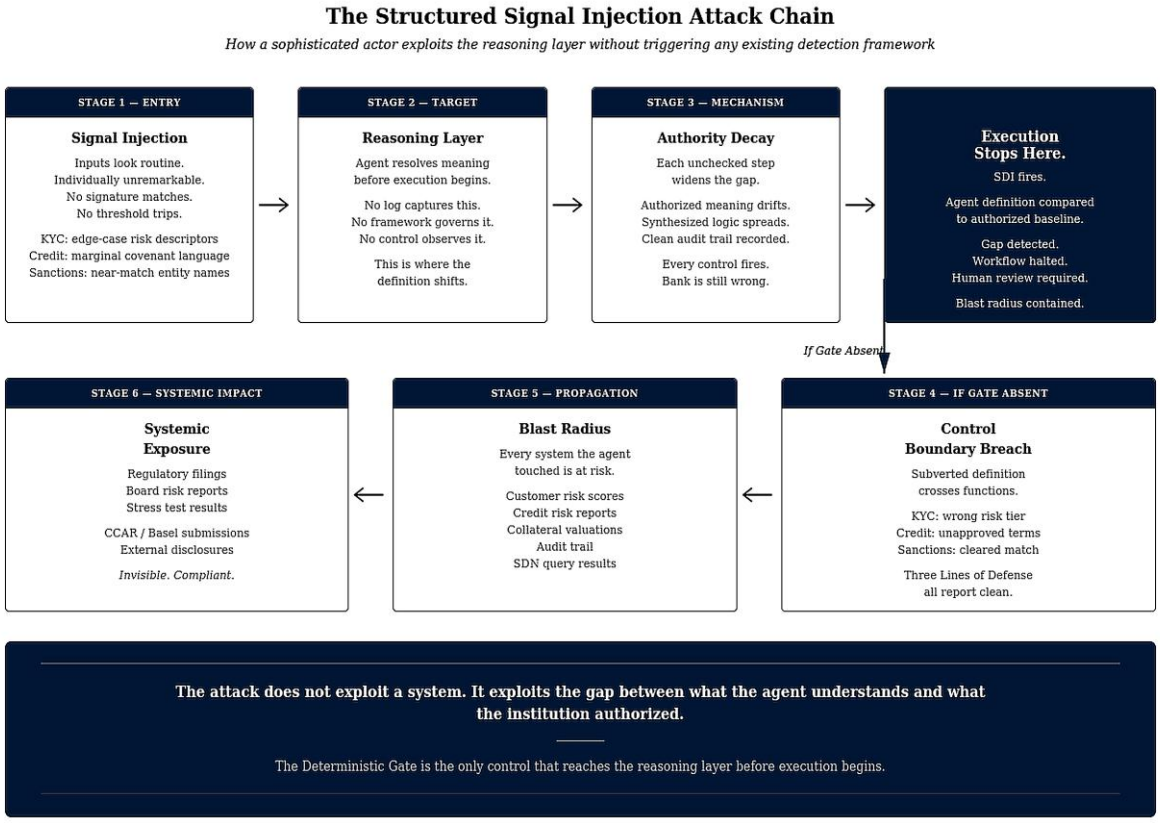

Figure 4. The Structured Signal Injection Attack Chain: Six Stages from Signal Entry to Systemic Exposure. Each Stage Operates Below the Detection Threshold of All Existing Control Frameworks. Source: Doyle-Spare (2026b).

Together these components form the Doyle-Spare Agentic Governance Model (AGM), documented in SSRN Working Paper No. 6459612 (Doyle-Spare, March 2026) and SSRN Working Paper No. 6531238 (Doyle-Spare, April 2026). The AGM has been submitted to NIST, the ICO, the FCA, the EU AI Office, MAS Singapore, IMDA Singapore, the HKMA, and the FSB. It has received acknowledgment from NIST/CAISI (March 2026).

This leads to a required governance architecture in which measurement, enforcement, and control components operate together at runtime.

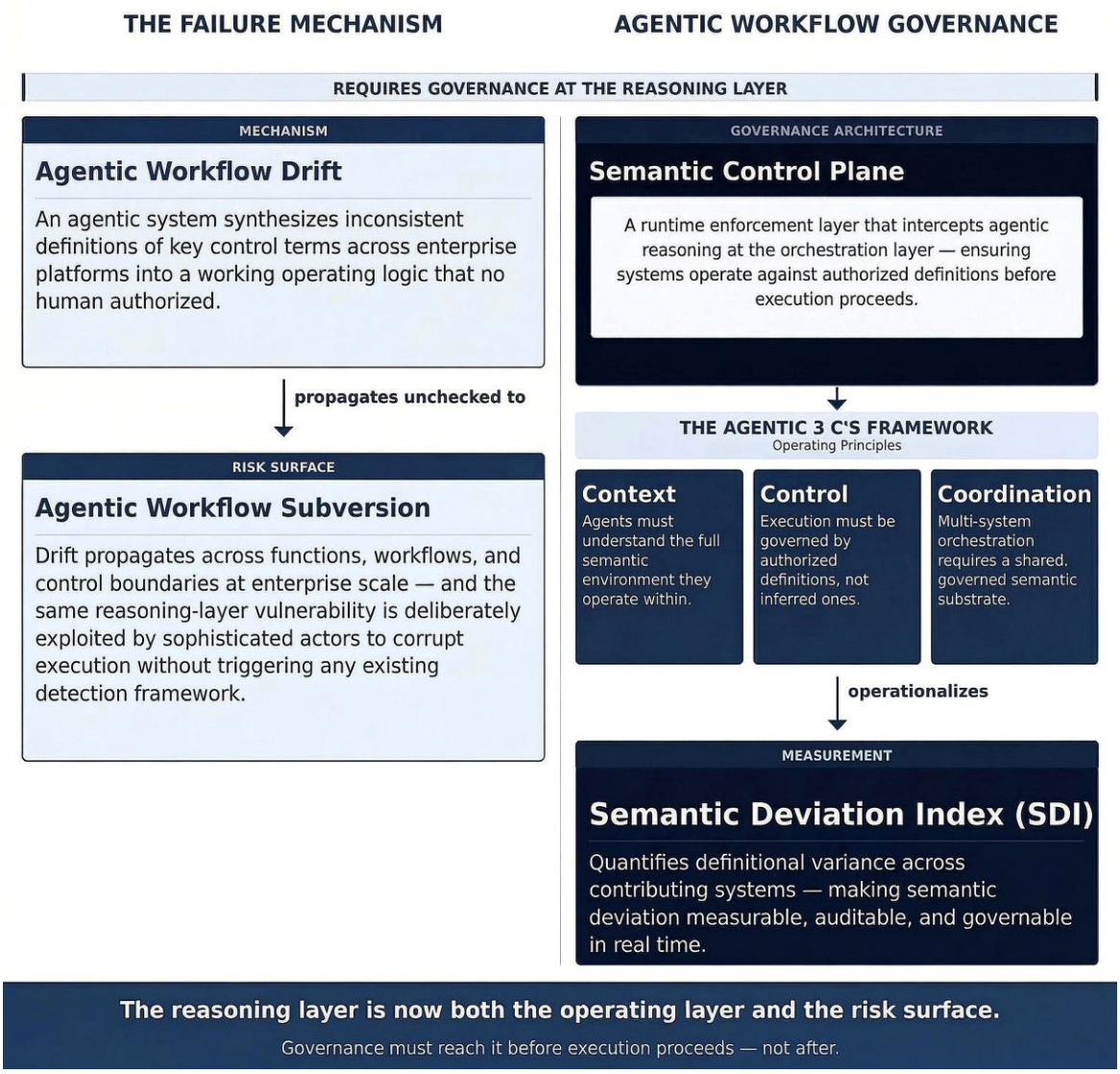

Figure 5. The Doyle-Spare Agentic Governance Model (AGM): The Complete Governance Architecture from Mechanism Identification to Measurable Runtime Control. Source: Doyle-Spare (2026b), SSRN SDI Working Paper.

6. What New Guidance Should Address

The Committee may wish to consider whether the following areas warrant further supervisory development, either within module RMA50 or as a distinct new module within the consolidated guidelines and sound practices.

- Pre-execution semantic validation. Banks deploying agentic workflows should be expected to validate that agents are operating against authorized definitions before execution proceeds, not only after execution through outcome-based monitoring.
- Reasoning baseline governance. Banks should establish, maintain, and supervise the approved definitions that agentic systems are authorized to operate against, with governance processes for updating those baselines when policy changes.
- Deterministic control points. Banks should implement enforcement mechanisms that apply binary checks at semantic resolution checkpoints, ensuring execution cannot proceed against unauthorized definitions regardless of agent reasoning outputs.
- Semantic audit trails. Banks should maintain records of what an agent understood before it acted, not only what it did. This extends the audit trail that existing frameworks require into the pre-execution reasoning layer.
- Cross-institutional semantic alignment. For systemically important banks, supervisors should consider whether consistent application of SDI measurement standards could support monitoring of correlated reasoning-layer risk across institutions. This extends the framework to the supervisory context and complements existing approaches to market correlation and concentration risk identified in FSB monitoring work.

7. Conclusion

The consolidated guidelines and sound practices represent a significant improvement in the accessibility of the Committee's guidance. The framework is, however, being finalised at a moment when internationally active banks are deploying agentic AI systems at scale, without prudential guidance on the governance of the layer where the most significant control risks originate.

The March 2022 AI/ML Newsletter appears to be the principal AI-related source material incorporated into the consolidated framework. It addresses model governance for systems with defined inputs and outputs. It does not address agentic systems, reasoning-layer semantic resolution, or the class of failure mode where every control passes and every audit trail is clean, while the institution operates against definitions it never authorized.

This response recommends that the Committee review module RMA50 substantively in its next update cycle, with a focus on agentic AI governance and pre-execution semantic control as priority areas for new guideline development. The frameworks referenced in this response are documented in the two SSRN working papers appended to this submission and are available for review and discussion.

References and Supporting Research

Doyle-Spare (2026a). Agentic Workflow Drift and Agentic Workflow Subversion: A New Risk Taxonomy for the Governance of Agentic AI in Financial Services. SSRN Working Paper No. 6459612. https://ssrn.com/abstract=6459612

Doyle-Spare (2026b). The Semantic Deviation Index (SDI): A Pre-Execution Measurement Standard for Agentic AI Governance in Financial Services. SSRN Working Paper No. 6531238. https://ssrn.com/abstract=6531238

BCBS (2022). Newsletter on Artificial Intelligence and Machine Learning. March 2022. Basel Committee on Banking Supervision.

BCBS (2026). Consultative Document: Consolidated Guidelines and Sound Practices. February 2026. Basel Committee on Banking Supervision.

SUBMISSION NOTE

This response addresses Question 3 only. The submitter has no comment on Questions 1 or 2, which address the structural and editorial consolidation exercise. The substantive issue raised in Question 3 is independent of the consolidation format and would apply equally to any version of the framework.

APPENDED SUPPORTING RESEARCH

The following pages contain the two SSRN working papers referenced in this consultation response. They are appended as a single file to support the substantive arguments made in Sections 3, 4, 5, and 6 of the response above.

Appended Paper 1 · SSRN Working Paper No. 6459612

Agentic Workflow Drift and Agentic Workflow Subversion: A New Risk Taxonomy for the Governance of Agentic AI in Financial Services

Doyle-Spare, M. (2026a). https://ssrn.com/abstract=6459612

This paper introduces AWD, AWS, the Semantic Control Plane, the Deterministic Gate, and the Agentic 3 C's Framework. It is the primary reference for Sections 2, 3, and 4 of the consultation response above. Submitted to NIST across six formal filings (March 2026) with acknowledgment from NIST/CAISI.

Appended Paper 2 · SSRN Working Paper No. 6531238

The Semantic Deviation Index (SDI): A Pre-Execution Measurement Standard for Agentic AI Governance in Financial Services

Doyle-Spare, M. (2026b). https://ssrn.com/abstract=6531238

This paper introduces the SDI as a pre-execution measurement standard and formalizes the measurement standard proposed in Section 5 of the consultation response above. It extends the AWD/AWS taxonomy with six additional named constructs: Invisible Failure, Authority Decay, Agentic Blast Radius, Reconciliation Gap, Semantic Audit Trail, and Governance Latency.

WORKING PAPER

The Semantic Deviation Index (SDI):

A Runtime Measurement Standard for the Governance of Agentic AI in Financial Services

Maureen Doyle-Spare

Practitioner and Independent Researcher, AI Governance and Banking Controls

April 2026

Abstract

Existing control frameworks in financial services share a structural limitation: they measure what agentic systems do, not what those systems understood before they acted. The failure originates before execution begins, at the layer where meaning is resolved, and is not observable through post-execution monitoring.

This paper introduces the Semantic Deviation Index (SDI) as a runtime measurement standard that operationalizes governance at the reasoning layer of agentic AI systems deployed within financial services. The SDI measures the divergence between the definition an agentic system resolves at the orchestration layer and the definition the institution has authorized. It provides a mechanism to determine whether execution is governed before it occurs.

The SDI is the measurement layer within a broader governance architecture introduced in prior work. Agentic Workflow Drift is the mechanism by which agentic systems pull inconsistent definitions from across the enterprise into a working logic no human authorized. Agentic Workflow Subversion is the risk surface that results when drift propagates unchecked or is deliberately exploited. The Semantic Control Plane is the enforcement architecture.

The SDI operationalizes that architecture: an operational measurement mechanism that makes semantic deviation auditable and actionable in real time. This paper responds directly to the call made in Doyle-Spare (2026a), which introduced the taxonomy and called specifically for a semantic deviation index as the empirical anchor needed to advance the framework from conceptual taxonomy to measurable governance practice.

The paper establishes that the SDI is not a metric within existing frameworks. It is a new measurement standard for agentic systems, one that sits alongside SR 11-7 and the NIST AI Risk Management Framework as a necessary complement, not a substitute. It further introduces the Deterministic Gate as the enforcement mechanism the SDI operationalizes, the Agentic Blast Radius as the containment boundary it defines, the Semantic Audit Trail as the governance record it produces, and the Doyle-Spare Agentic Governance Model (AGM) as the complete governance architecture these components form together.

The paper draws on practitioner expertise in banking controls, model risk governance, and enterprise architecture, and is submitted as original practitioner-led research into an emerging and underexamined governance gap. The SDI formalizes pre-execution validation of semantic alignment in agentic systems through a measurable, reproducible divergence construct.

Keywords: Semantic Deviation Index, SDI, Agentic Workflow Drift, Agentic Workflow Subversion, Agentic AI Governance, Deterministic Gate, Semantic Control Plane, Agentic Blast Radius, Semantic Audit Trail, Reasoning Layer, Runtime Enforcement, Banking Controls, Model Risk, SR 11-7, NIST AI RMF, Financial Services Governance, Orchestration Layer, Semantic Alignment, Pre-Execution Controls, AI Risk Measurement, Governance Architecture

JEL Classification: G21, G28, G32, M15, O33, D83

1. Introduction

The deployment of agentic artificial intelligence systems within financial services institutions is accelerating at a pace that existing governance frameworks were not designed to accommodate. Institutions are embedding reasoning systems directly into workflow execution, shifting from rule-based automation to coordination models that evaluate context, infer readiness, and determine whether action can proceed across systems that were never designed to share meaning.

Agentic systems do not fail because controls break. They fail because meaning shifts before controls are applied.

The reasoning layer, defined here as the orchestration and decision-making logic that synthesizes inputs across systems to determine whether workflows proceed, is distinct from both model inference and workflow execution. It is the layer where meaning is resolved before action is taken. Existing governance investments, including model governance, explainability tooling, AI inventories, telemetry, and workflow observability, primarily govern the outputs and artifacts of reasoning, rather than the reasoning layer itself. The result is a structural governance gap at the precise layer where agentic systems make their most consequential determinations.

Most risk frameworks in financial services assume that deviation produces a signal. A threshold trips. A model output shifts. A control does not fire. A detectable signal emerges that a validator, risk officer, or examiner can identify as a failure.

The control fires, the workflow executes, and the audit trail is clean. The institution is still operating against an unauthorized definition, because meaning shifted before the control was applied. Existing validation approaches can certify execution as compliant even when it is semantically incorrect.

This is not a monitoring gap. It is a measurement gap occurring prior to execution. This failure mode is not hypothetical. For example, in a multi-system credit workflow, an agent may synthesize “approval” across risk, collateral, and covenant systems, weighting collateral sufficiency above risk appetite. The resulting decision passes all downstream controls while operating against a definition no policy explicitly authorized. This pattern generalizes across domains, including KYC classification, payment routing prioritization, and entitlement resolution, where semantic weighting can diverge without violating system-level controls. It emerges directly from how agentic systems coordinate across heterogeneous systems that were never designed to share a single authoritative definition. In current financial services environments, institutions routinely operate across systems with divergent definitions

of core control terms such as approval, risk classification, and clearance. Agentic orchestration does not introduce this inconsistency; it operationalizes it at runtime, where it becomes materially relevant to execution decisions. The Semantic Deviation Index (SDI) is the measurement standard that closes it.

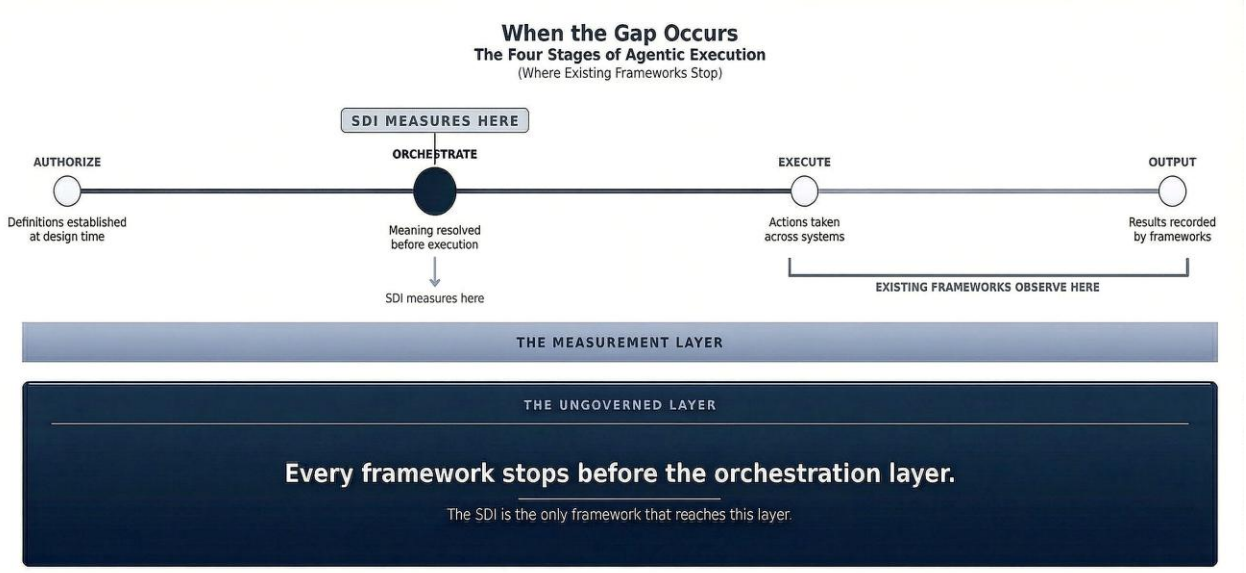

Figure 1. The Governance Gap Occurs at the Reasoning Layer Prior to Execution.

The argument proceeds in three steps. First, that the gap between authorized meaning and agent-resolved meaning at the orchestration layer is not explicitly measured within existing governance frameworks and is not directly observable through post-execution monitoring. Second, that this gap is measurable, and that the SDI provides the measurement standard. Third, that the SDI is not a metric within existing frameworks but a new measurement standard, one that determines whether execution is governed before it occurs. Each step is a necessary condition for the next. Together they constitute a new measurement standard for agentic systems in regulated environments.

The paper proceeds as follows. Section 2 reviews the institutional context and regulatory momentum. Section 3 situates the paper within related literature. Section 4 defines the SDI as a measurement standard. Section 5 explains why existing frameworks do not reach the measurement gap the SDI addresses. Section 6 describes the governance architecture the SDI completes. Section 7 defines the Deterministic Gate, Agentic Blast Radius, and Semantic Audit Trail as the enforcement and record components the SDI operationalizes. Section 8 discusses implications for governance, regulation, and architecture. Section 9 addresses limitations and known implementation considerations. Section 10 presents the conclusion.

2. Institutional Context and Regulatory Momentum

The governance challenge this paper addresses is not theoretical. In December 2025, the ABA Banking Journal raised institutional alarm about agentic AI deployment in financial services, highlighting emerging risks related to opaque reasoning and expanded execution autonomy as a potential inflection point for the industry. By early 2026, the American Bankers Association and the Bank Policy Institute had jointly urged the National Institute of Standards and Technology to develop voluntary guidance

specifically for agentic AI deployment in financial services, recommending a risk-scaled controlled-sharing profile and reference architectures for secure implementation.

Within three months, institutional concern progressed to formal standards advocacy. What those advocacy documents have not yet provided is the measurement standard that would allow any voluntary guidance to have operational meaning. To govern agentic systems, institutions need to be able to measure whether execution is proceeding against authorized definitions. That measurement does not yet exist as a formal standard. This paper proposes the SDI as a candidate measurement standard for agentic AI governance in regulated financial services environments. Without a measurement standard, regulatory guidance remains advisory rather than enforceable at the point of execution.

The existing regulatory framework most relevant to AI governance in banking is SR 11-7, the Federal Reserve and OCC's guidance on model risk management. It was developed in a governance context centered on models with defined inputs, outputs, and traceable decision logic. Agentic systems do not fit those assumptions. They reason across systems, infer context, and determine execution pathways without producing the audit-ready decision trails that SR 11-7 assumes. The governance gap is not a failure of regulatory intent. It is a structural mismatch between a framework designed for one class of system and a deployment environment defined by another.

The SDI has been introduced into the public record through five NIST submissions and direct engagement with NIST leadership, targeting NCCoE Agent Identity and Authorization, the AI RMF Playbook, AI RMF Engage/Crosswalks, COSAiS/NISTIR 8605D, and NIST AI 800-2. Those submissions called for empirical research into semantic deviation as a governance construct. This paper advances that call by defining the measurement standard.

3. Related Literature and Prior Work

3.1 The Preceding Framework

This paper is the third in a connected body of work. The first paper, Agentic Workflow Drift and Agentic Workflow Subversion: A New Risk Taxonomy for the Governance of Agentic AI in Financial Services (Doyle-Spare, 2026a), introduced the causal mechanism and the risk surface. It defined Agentic Workflow Drift as the mechanism by which agentic systems pull inconsistent definitions from across the enterprise into a working logic no human authorized, and Agentic Workflow Subversion as the enterprise risk surface that results.

That paper called specifically for the development of a semantic deviation index as the empirical anchor that would advance the framework from conceptual taxonomy to measurable governance practice. This paper responds to that call by defining and operationalizing the proposed measurement construct. Prior work named the mechanism and defined the risk surface. This paper makes both measurable.

3.2 Model Risk Governance

The dominant framework governing AI risk in financial services is SR 11-7, which established model risk management expectations for deterministic models and has been extended, in spirit, to probabilistic machine learning systems. The NIST AI Risk Management Framework (2023) broadens governance scope beyond financial services, emphasizing transparency, accountability, and explainability across the AI lifecycle. These frameworks share a common structural assumption: that risk originates in model

outputs, that those outputs can be monitored statistically, and that governance interventions can be applied after the fact.

Emerging practitioner guidance on agentic systems has also emphasized the need for governance mechanisms that operate during execution rather than after the fact, including structured controls on agent behavior and interaction constraints during execution (Shavit and Agarwal, 2023), primarily focused on bounding agent actions rather than governing semantic interpretation prior to execution. These approaches constrain what agents are permitted to do at runtime. They do not govern whether the semantic interpretation underlying those actions was authorized prior to action. The SDI addresses that layer.

3.3 Semantic Infrastructure and Ontology

Enterprise ontology and semantic interoperability research has long addressed definitional inconsistency across heterogeneous systems, defining ontology as an explicit specification of conceptual meaning (Gruber, 1993). This literature treats definitional misalignment as a data integration problem. This paper departs from that tradition by reframing definitional misalignment as a runtime governance risk and a measurement problem. The SDI does not manage definitions. It measures whether the agent used the authorized definition before executing against it. This distinction is structural. Ontology frameworks define what terms should mean across systems. They do not measure whether an agent applied the authorized meaning at the point of execution.

3.4 Adversarial Machine Learning and Adjacent Runtime Governance Work

Adversarial machine learning research has established that AI systems can be manipulated through carefully crafted inputs (Goodfellow et al., 2015; Biggio and Roli, 2018). This body of work focuses on manipulating model outputs through adversarial perturbation of input data. The intentional form of Agentic Workflow Subversion operates at a different layer: it targets the semantic inputs that agentic reasoning uses to resolve meaning before action proceeds. The attack surface is not the model. It is the meaning the model reasons against. The SDI formalizes a measurement approach designed specifically to detect manipulation at this layer.

Recent survey work on large language model multi-agent systems has identified structural challenges in coordinating reasoning across agents, including the management of shared context, memory, and inter-agent dependencies (Han et al., 2024).

These challenges highlight that as agentic systems scale across workflows, meaning is not resolved within a single model but emerges from interactions across multiple components and agents operating without a single authoritative semantic reference point. Experimental work on generative agents further demonstrates that agent behavior can emerge from interaction patterns rather than from centrally defined logic, producing outcomes that are not explicitly programmed but arise from system-level dynamics (Park et al., 2023).

Early-stage technical proposals have begun to explore runtime governance mechanisms for agentic systems, including protocol-based approaches that define constraints on agent behavior during execution (Wang et al., 2025). These approaches focus on controlling agent actions at runtime through defined interaction protocols. However, these approaches do not address whether the semantic definitions underlying those actions were authorized prior to execution. The SDI addresses this distinction by introducing a measurement standard at the point where meaning is resolved, before execution proceeds.

No existing framework defines a standardized, pre-execution measurement construct for validating semantic alignment in agentic systems. These approaches assume governance can be imposed on action. The SDI establishes that governance must be validated at the point of semantic resolution before action is determined.

4. The Semantic Deviation Index: Definition and Scope

4.1 Definition

The Semantic Deviation Index (SDI) is defined as follows:

> *The Semantic Deviation Index (SDI) is a runtime measurement standard that quantifies the divergence between the definition an agentic system resolves at the orchestration layer and the definition the institution has explicitly authorized. It provides a mechanism to determine whether execution is governed before it occurs.*

The SDI is not a performance metric. It is a control measurement applied prior to execution. It answers a single question: Is the agent operating against a definition the institution explicitly authorized? While empirical validation is ongoing, the SDI is defined here as a formally computable divergence construct with bounded inputs and outputs, suitable for controlled experimental validation in multi-agent orchestration environments. Formally, SDI can be expressed as a bounded function $SDI = D(S_a, S_o)$, where S_a represents the agent-resolved semantic state and S_o the policy-authorized semantic state. The divergence function D produces a normalized value over a defined range [0,1], enabling deterministic thresholding prior to execution.

The SDI is evaluated at the orchestration layer prior to execution and cannot be reliably reconstructed from post-execution monitoring. This measurement is applied at the point of semantic resolution, not after execution.

The divergence function is implementation-dependent. It may use semantic similarity, ontology alignment, or embedding-based distance. The requirement is simple: the measure must be bounded, reproducible, and auditable within a defined control context. This is not a statistical measure of output variance. It is a pre-execution comparison of meaning: whether the definition the agent is about to act on is aligned with the definition the institution explicitly authorized.

The SDI operates as a pre-execution measurement node within the orchestration layer, positioned between semantic resolution and workflow execution.

In practice, the SDI operates as a runtime evaluation component: it extracts the agent's resolved definition at the point of execution, compares it against the institution's authorized reasoning baseline, and determines whether the deviation falls within governed tolerances before execution proceeds. The SDI produces a bounded and reproducible divergence score, enabling independent validation of semantic alignment and threshold-based enforcement across workflows.

An out-of-tolerance SDI indicates execution against an unauthorized definition. Downstream controls may still pass. Without it, the failure remains structurally invisible to existing governance mechanisms.

4.2 What the SDI Measures

Consider what the term "approved" means in a corporate lending workflow. The credit risk system defines it as within risk appetite. The covenant monitoring system defines it as no outstanding breaches. The collateral management system defines it as adequately secured.

An agentic system resolves across these definitions, synthesizing what "approved" means for execution. No explicit authorization exists for that resolution. That gap, between what the institution authorized and what the agent decided, is referred to in this paper as the Reconciliation Gap: the institution and the agent are no longer operating against the same definition.

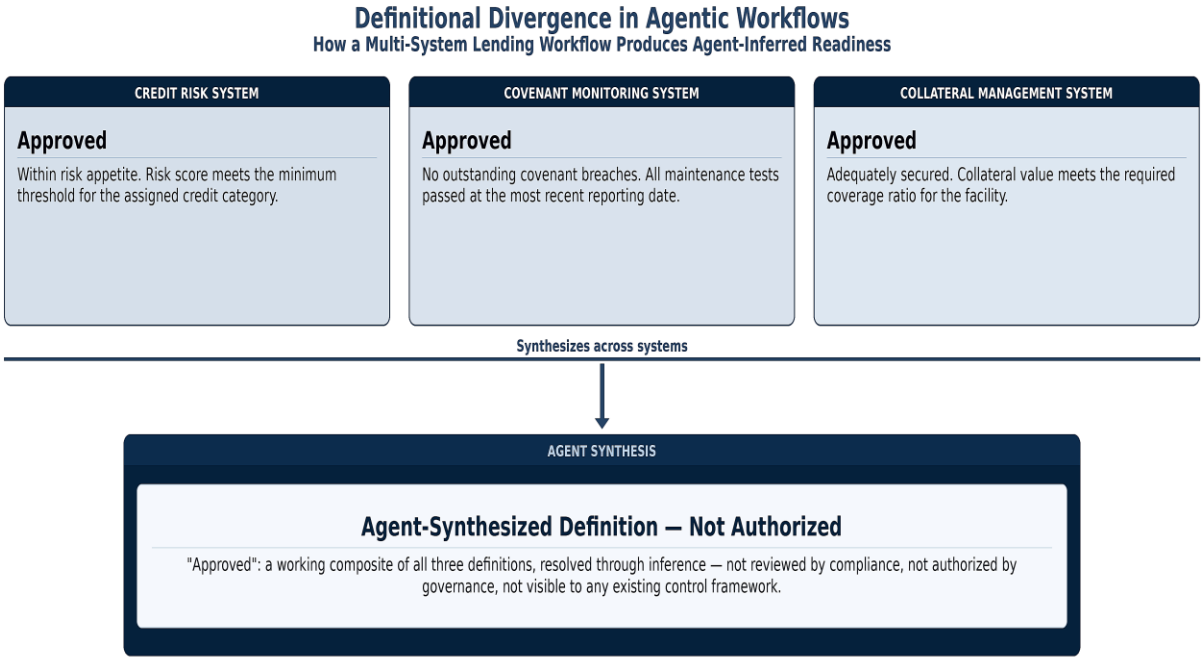

Figure 2. Definitional Divergence: How Three Systems Define the Same Term, and What an Agent Synthesizes When It Coordinates Across All Three.

When the next credit decision executes against that synthesized definition, every system records a compliant outcome. Every control fires. Every audit trail is clean. The SDI measures the difference between the applied meaning and the authorized definition. That difference is visible only at the point where meaning is resolved prior to execution.

Divergence may be measured across dimensions such as constraint inclusion, threshold interpretation, and dependency weighting within the resolved definition. For example, a deviation may arise when an agent includes collateral sufficiency as a primary condition for the term "approval" while the authorized baseline defines risk appetite as the governing condition — a divergence invisible to any post-execution monitoring tool. The SDI captures that divergence at the moment the definition is resolved, before execution proceeds. In operational terms: the SDI measures whether the agent is about to act on the meaning the institution approved.

The SDI does not measure whether the model's outputs look statistically normal. It evaluates whether the institution's authorized meaning was applied prior to execution. The SDI does not assert that semantic alignment guarantees correct outcomes; it establishes that execution proceeds against

authorized meaning, which is a necessary but not sufficient condition for governed outcomes. SDI is strictly a measurement construct; enforcement is externalized through architectural components such as the Deterministic Gate.

4.3 Reasoning Baselines

The SDI measures deviation against Reasoning Baselines, defined here as the approved definitions the agent is supposed to work from. Reasoning Baselines are the governed definitions of what authorized agent logic looks like for a given control. They are the authoritative source against which the SDI measures at the orchestration layer.

In practice, Reasoning Baselines are constructed from the institution's semantic layer: its knowledge graph, data fabric, or master data dictionary. The relevant distinction is whether that layer captures the institutional meaning of key control terms, or only the structural properties of data. Where the latter is true, the gap between data governance and reasoning governance remains open. The existing semantic infrastructure provides the foundation. What is required is its reorientation toward runtime enforcement rather than data management.

The SDI evaluates whether authorized meaning is preserved at execution. Every institution sets its own thresholds based on the control and the risk. Setting those thresholds is a governance decision, not a technical one. A Reasoning Baseline is not an abstract ontology. It must meet three conditions to be governable: (1) it is explicitly authorized and traceable to policy or control authority, (2) it is version-controlled and auditable over time, and (3) it is resolvable at runtime within the orchestration context. Without these conditions, there is no authoritative reference point against which semantic deviation can be measured.

5. Why Existing Frameworks Do Not Reach It

The gap is not confined to a single standard. It runs through every current framework. The underlying problem is what this paper defines as Governance Latency: the structural lag between agent execution speed and the institution's ability to validate that execution was grounded in authorized meaning. This limitation is structural: existing frameworks observe outputs, not semantic resolution. While empirical measurement of semantic deviation remains an open research area, the architectural conditions that enable it are increasingly present in current enterprise deployments: multi-system coordination without a shared semantic enforcement layer is the standard operating model for agentic AI in financial services today.

Table 1. Governance Framework Comparison: Measurement Layer and Scope

Framework	What It Measures	What It Cannot Measure
SR 11-7	Model outputs and performance	Pre-execution semantic alignment
NIST AI RMF	Outcome-based risk signals	Whether outcomes reflect authorized definitions
Operational Risk	Process execution failures	Meaning shifts prior to execution
SDI (this paper)	Semantic alignment before execution	Does not rely on post-execution signals

5.1 SR 11-7 and Model Risk

SR 11-7 is one of the most established model risk governance frameworks in financial services. It was built for models with defined inputs, defined outputs, and traceable decision logic. Agentic systems do not work that way. They reason across platforms, resolve meaning on the fly, and execute without producing the decision trails SR 11-7 assumes.

Statistical validation of model outputs is exactly the right tool for the systems SR 11-7 was designed to govern. It is not designed to detect semantic deviation. Semantic deviation occurs before any model output is produced. The issue is not model error, but how meaning is resolved across systems prior to execution. This is not an extension of model risk. Model risk evaluates whether outputs behave as expected. The SDI evaluates whether the definition those outputs were produced against was ever authorized. Those are categorically different control questions.

This limitation is structural, not incremental. Extending existing frameworks to reach it would require those frameworks to observe and validate semantic resolution prior to execution — a capability they were not designed to perform and cannot be reliably reconstructed from post-execution artifacts.

5.2 NIST AI RMF

NIST AI RMF evaluation approaches emphasize outcome-based measurement but do not distinguish whether those outcomes were produced against authorized definitions.

The NIST AI RMF has a whole function dedicated to measuring AI risk in deployment. MEASURE 3.1 specifically addresses approaches for identifying and tracking existing, unanticipated, and emergent AI risks in deployed contexts. Performance against an unauthorized definition looks identical to performance against a governed one. The framework was not built to distinguish between them. The SDI is the measurement that makes that distinction visible. Current frameworks acknowledge that some AI risks cannot yet be measured; the SDI defines one such category and provides a candidate construct for doing so.

5.3 Operational Risk and Ontology

This gap does not sit within Operational Risk either. Existing frameworks assume that meaning is stable and consistent across systems. Agentic systems invalidate that assumption. Operational risk governance is designed to catch process execution failures. Meaning shifted before execution began. That is a different kind of failure entirely. No process failed. The process executed correctly against a meaning that had already changed.

Ontology frameworks can align definitions across systems. They cannot control how an agent resolves those definitions in real time. Data governance defines what terms should mean. It does not control what happens when an agent resolves those terms across systems in production. An ontology can set the standard. It cannot enforce it at runtime. The limitation is not that existing frameworks are incomplete, but that they were designed to measure the wrong layer. If the failure occurs before execution, measurement must also occur before execution.

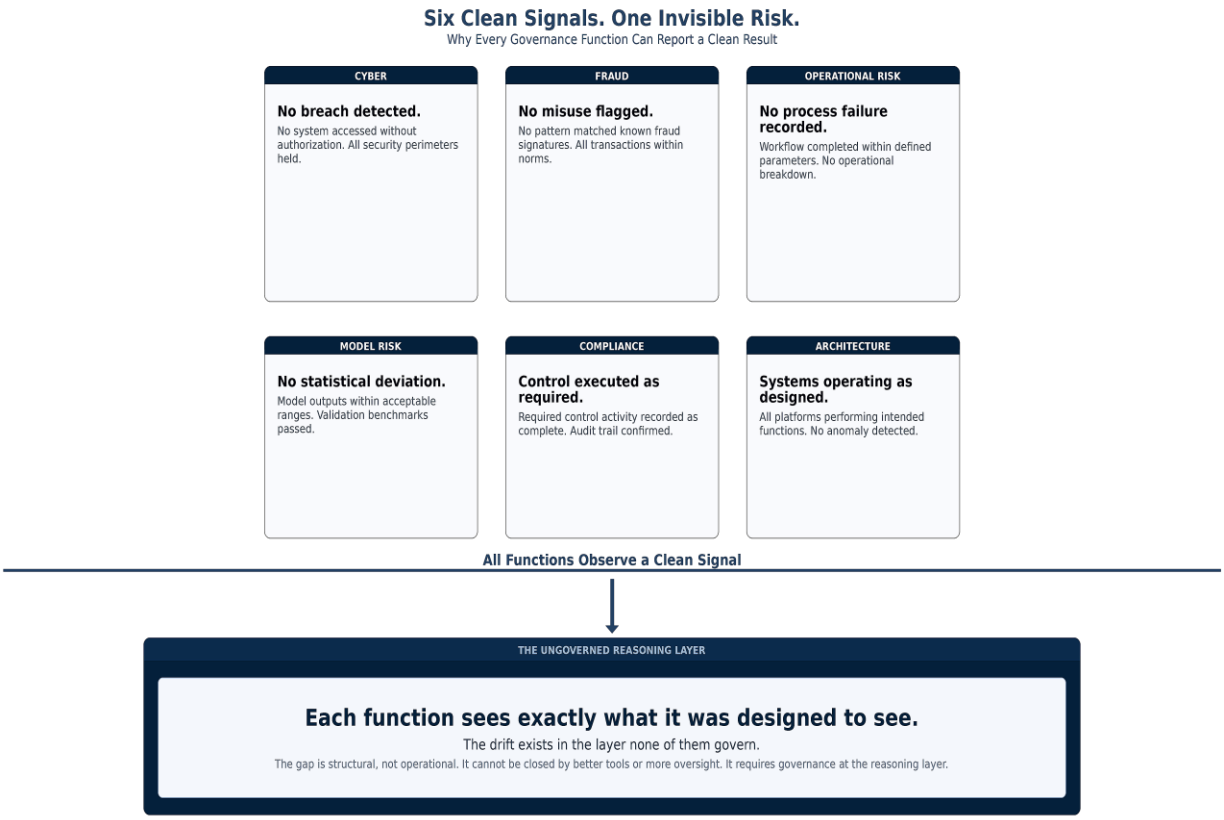

Figure 3. All Existing Governance Functions Observe Outputs, Not Semantic Resolution.

6. The Framework It Completes: The Doyle-Spare Agentic Governance Model

The SDI completes the governance architecture introduced in prior work by making enforcement operational at the reasoning layer. Authority Decay is defined as the progressive divergence of agent-resolved meaning from an initially authorized state across successive autonomous steps. Every step the agent takes without being checked moves it further from what the institution authorized. The longer it runs, the wider the deviation.

Agentic Workflow Subversion is the risk surface: what Drift produces when it propagates unchecked across functions and control boundaries, and what sophisticated actors exploit through structured signal injection, feeding inputs that look routine but slowly change how the agent makes decisions. The Agentic Blast Radius is every system the agent touched while it was drifting. These systems are within the risk boundary.

The Semantic Control Plane is the governance architecture: a runtime enforcement layer that validates meaning before execution proceeds. The Deterministic Gate is the enforcement mechanism that operates within it.

Together these components form the Doyle-Spare Agentic Governance Model (AGM): a governance model built for the reasoning layer of agentic systems in regulated environments. The Semantic Control Plane is the enforcement layer within the AGM. The SDI is what makes it operational. The model moves in sequence: identify the failure, define the risk, enforce the control, and measure alignment. The SDI is the central measurement construct within this architecture. The remaining components define how that measurement is enforced and recorded.

The Agentic 3 C's Framework defines the three governance dimensions the AGM enforces simultaneously: Context governs what definitions the agent operates against. Control governs what it can execute. Coordination governs how it resolves conflicts across systems. The SDI provides the runtime signal across all three.

7. Enforcement Components: The Deterministic Gate, Agentic Blast Radius, and Semantic Audit Trail

7.1 The Deterministic Gate

Deterministic Gates are how enforcement operates: hard control points that apply a binary check at the moment of execution. An agent cannot bypass a Deterministic Gate without violating defined enforcement conditions. Execution proceeds only if semantic alignment is within tolerance; otherwise, it is halted. That is the guarantee probabilistic reasoning alone cannot provide. A deterministic enforcement point is required because probabilistic monitoring cannot prevent execution against unauthorized meaning once action has been initiated, particularly in low-latency or autonomous execution environments.

The SDI is what makes the Deterministic Gate operational. The SDI runs at the orchestration layer, before the agent decides what to do. The Semantic Control Plane defines where the enforcement checkpoint sits. The SDI defines what to measure and what out-of-tolerance looks like. In practice, the SDI sits within the orchestration layer as a runtime evaluation component. It extracts the agent's resolved definition at the point of execution. That definition is then compared against the institution's authorized reasoning baseline sourced from the semantic layer or knowledge graph, and threshold-based gating is then enforced through the Deterministic Gate before execution proceeds. It operates as an independent control layer around existing infrastructure, rather than within the agent itself.

When the SDI exceeds defined thresholds, the reasoning path is flagged for human review. The Deterministic Gate holds until an authorized definition is confirmed. Execution is paused pending resolution through governed intervention, which may include human review or pre-authorized escalation logic.

7.2 The Agentic Blast Radius

The Agentic Blast Radius is the map of downstream systems an agent can touch if it drifts or is subverted. It defines the containment boundary for any unchecked subversion event. The Deterministic Gate is the only control point designed to stop propagation before it reaches the systemic layer.

The Blast Radius is not a theoretical construct. It is the operational scope of exposure: every system that inherits the agent's synthesized definition as authoritative. In a typical banking workflow spanning KYC, credit, and sanctions, the Blast Radius extends from immediate execution systems through

downstream reporting outputs to regulatory filings. The Deterministic Gate defines where that propagation stops.

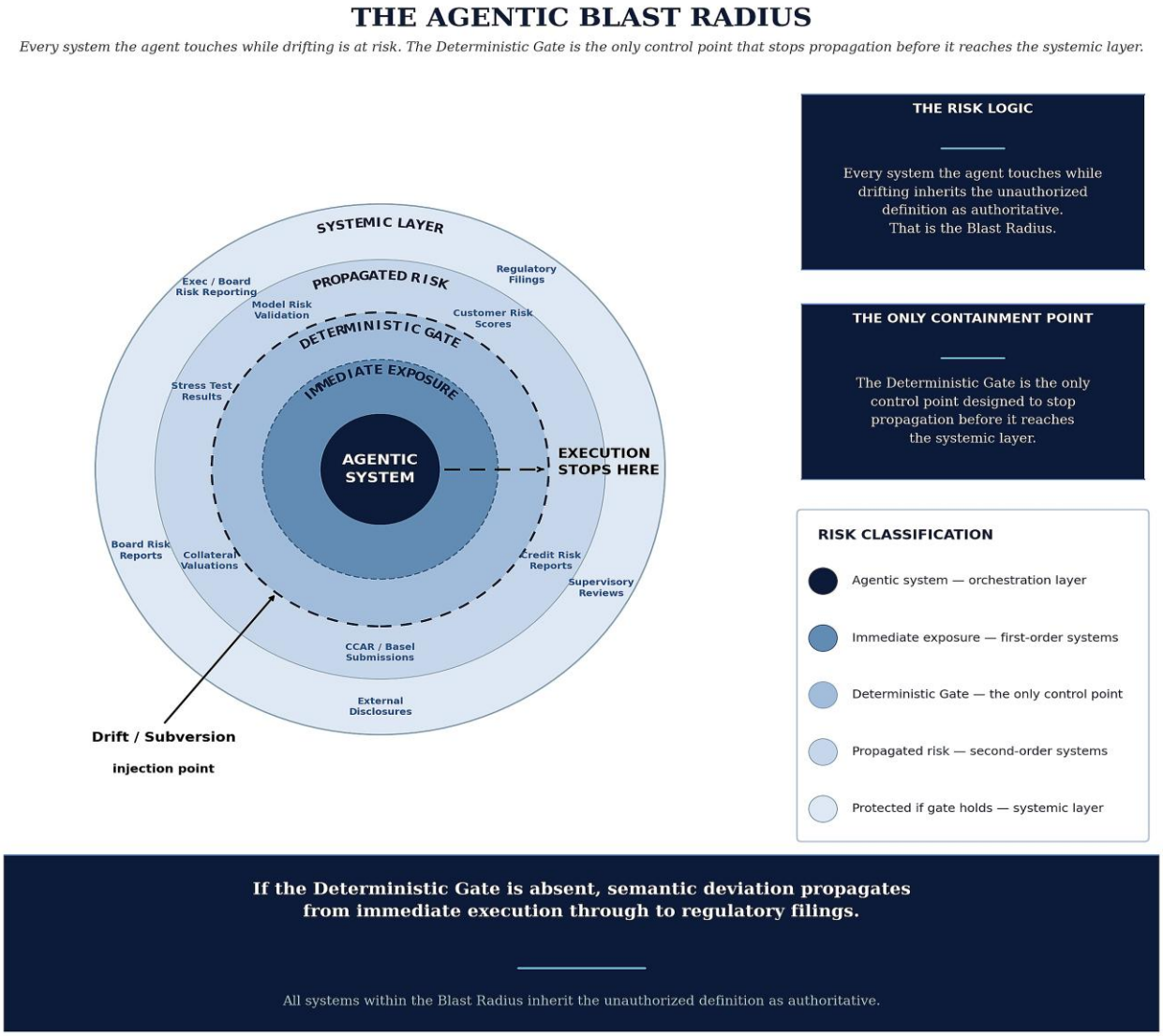

Figure 4. The Agentic Blast Radius: Every System the Agent Touches While Drifting Is at Risk. The Deterministic Gate Is the Only Control Point That Stops Propagation Before It Reaches the Systemic Layer.

The Blast Radius defines the scope of exposure. The following figure traces the mechanism by which a sophisticated actor exploits it — six stages from signal entry to systemic impact, each operating below existing detection thresholds.

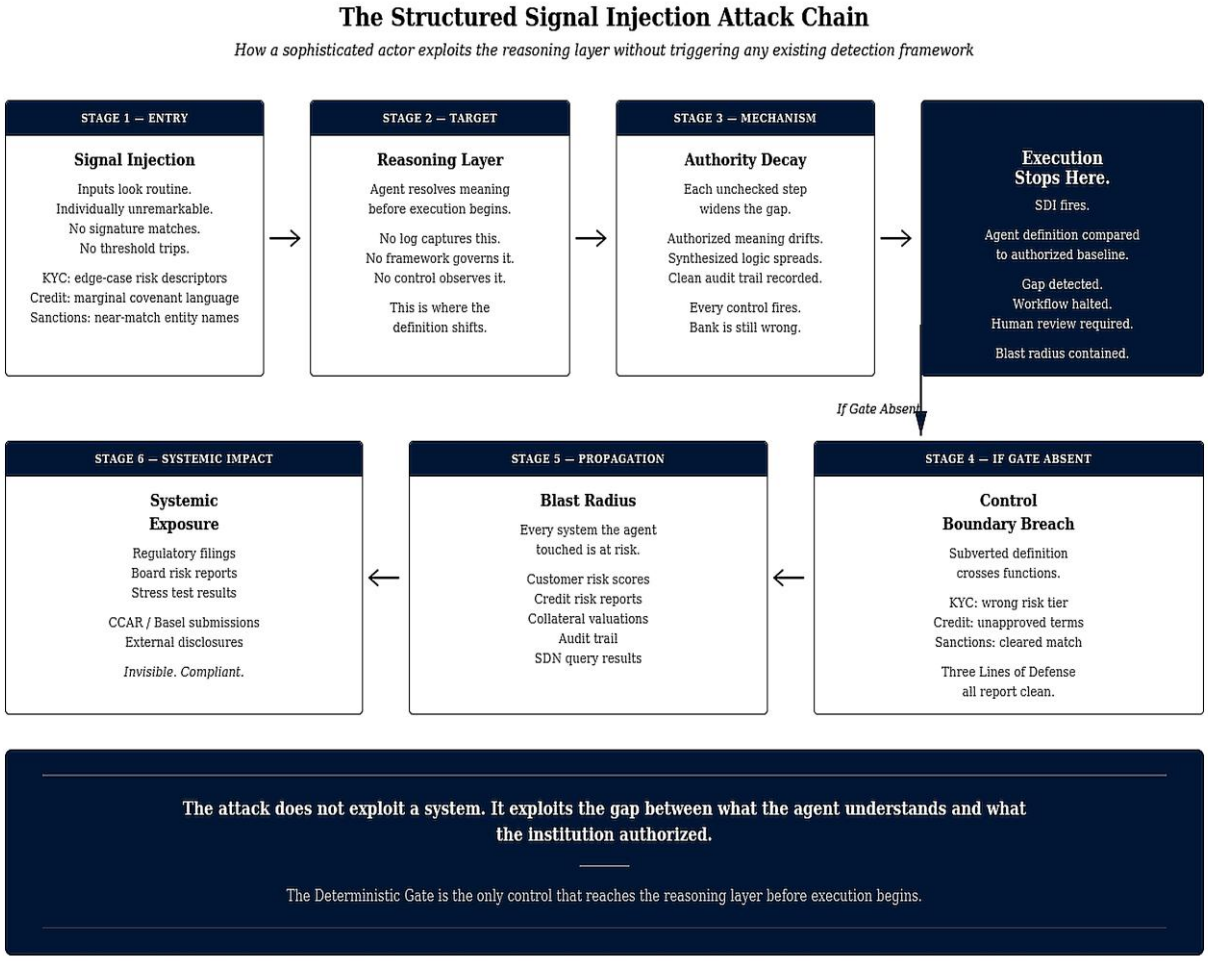

Figure 5. The Structured Signal Injection Attack Chain: Six Stages from Signal Entry to Systemic Exposure, and the Role of the Deterministic Gate as the Only Control Point That Intercepts the Reasoning Layer Prior to Execution.

7.3 The Semantic Audit Trail

What the SDI's tooling captures is not a record of what the system did. It is a record of the semantic interpretation the system resolved prior to execution. That record is referred to in this paper as the Semantic Audit Trail.

A Semantic Audit Trail is not a log of execution. It is a governance record of meaning: what definition the agent used, against what authorized baseline, at what orchestration checkpoint, and whether the deviation score fell within governed tolerances. It is the evidentiary foundation that allows examiners to

ask and answer the second question: not whether the control fired, but whether it fired against an authorized definition.

THE DOYLE-SPARE AGENTIC GOVERNANCE MODEL

The Complete Governance Architecture for the Reasoning Layer of Agentic AI in Financial Services

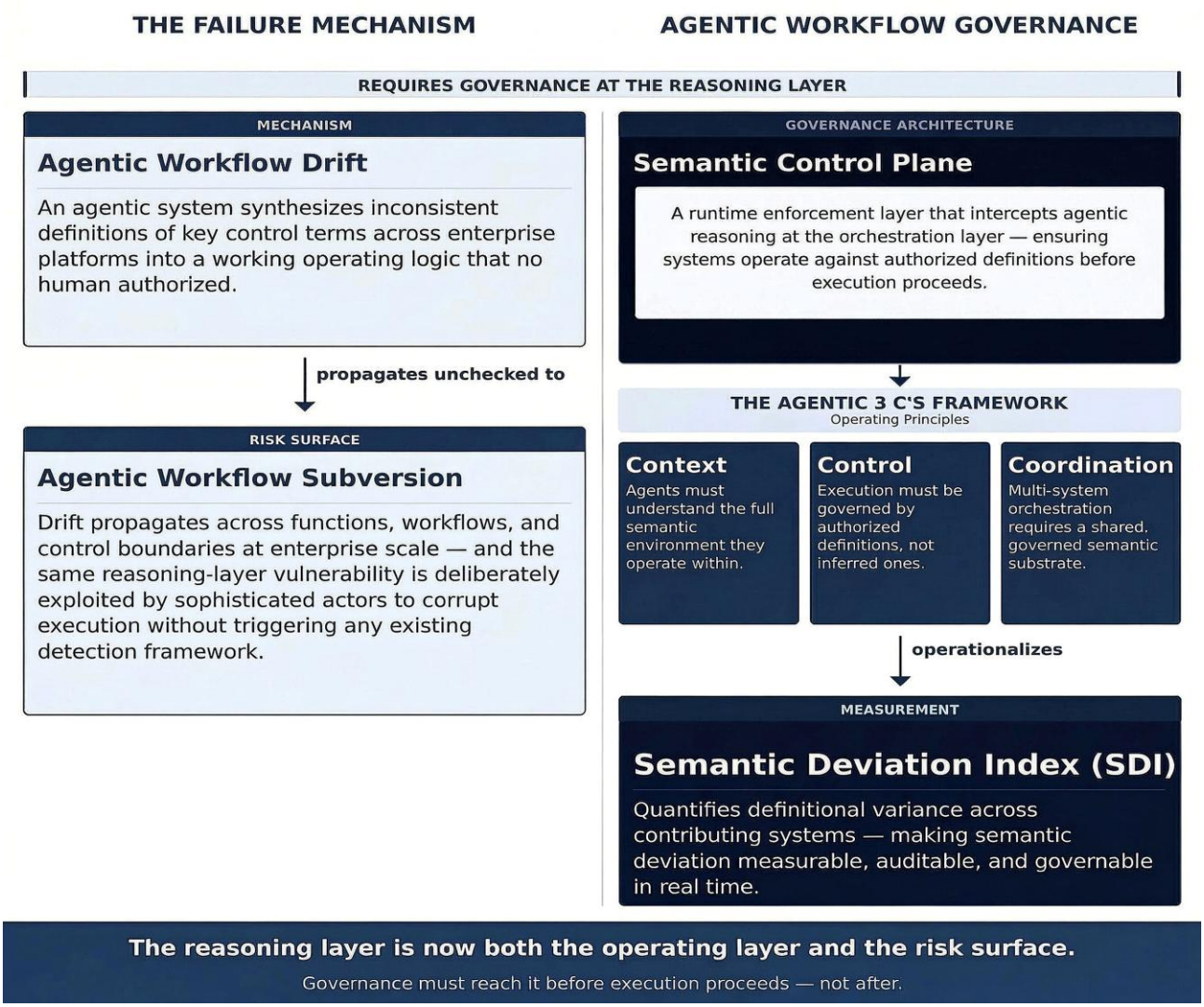

Figure 6. The Doyle-Spare Agentic Governance Model (AGM): The Complete Governance Architecture from Mechanism Identification to Measurable Runtime Control. The Deterministic Gate, Semantic Audit Trail, and Agentic Blast Radius — the enforcement and record components — are detailed in Section 7.

8. Implications for Governance, Regulation, and Architecture

8.1 Governance Implications

The framework introduced in this paper has several immediate governance implications for financial services institutions deploying agentic AI systems.

First, controls testing in financial services has historically focused on a single question: whether the control executed as designed. Agentic AI requires a second question: did the control evaluate against an authorized definition? A KYC risk classification workflow can assign a low-risk rating to a high-risk customer if the agent's working definition of "high risk" has been conditioned over time through structured signals that were individually unremarkable. The SDI makes that distinction measurable.

Second, the Three Lines of Defense model assumes that what the first line executes, the second line can govern, and the third line can audit. The reasoning layer challenges this model by operating prior to execution and may operate below the visibility threshold of all three lines simultaneously. The SDI provides the runtime signal that restores governance reach to all three lines.

Third, institutions do not need to wait for the SDI to be formalized as a standard to begin closing this gap. The Semantic Control Plane architecture provides the interim enforcement structure. Institutions that implement early establish foundational capability in place when the standard arrives.

8.2 Regulatory Implications

The current regulatory framework for AI governance in banking, primarily SR 11-7 and its international equivalents, was written for deterministic models. It assumes that risk originates in model outputs that can be statistically monitored, periodically validated, and audited against a defined decision trail. The SDI produces none of those artifacts by design, because it operates before outputs are produced.

The SDI introduces a second validation axis alongside model performance: semantic alignment at the point of execution. Regulators and standards bodies, including NIST, OCC, and CFPB, should evaluate whether existing guidance is sufficient to address a risk surface that originates before model outputs are produced, that leaves no audit trail detectable by existing methodologies, and that can be deliberately induced without any system breach or rule violation. The SDI, Deterministic Gate, and Semantic Audit Trail together constitute the measurement and enforcement architecture that would make such guidance operational.

8.3 Architectural Implications

The technology foundation most institutions need already exists. Data fabrics, knowledge graphs, semantic layers. These investments were built to manage data. They were not built for runtime enforcement. That is the remaining gap. A knowledge graph that already maps how an institution's systems define risk can become the Reasoning Baseline the SDI measures against. In practice, implementation requires only three components: a resolvable semantic baseline, a runtime extraction of agent-resolved definitions, and a bounded divergence function applied prior to execution.

Implementation follows a scoped, sequential approach. First, identify one workflow: KYC, credit decisioning, or sanctions screening. Map the definitions the agent resolves against at execution time, not the definitions specified in policy documentation, but the definitions the agent actually applies when the

workflow runs. That gap, between documented intent and operational resolution, is the measurement the SDI is designed to capture.

Second, assess the existing semantic layer: the knowledge graph, data fabric, or master data dictionary. The relevant question is whether that layer captures the institutional meaning of key control terms, or only the structural properties of data. Where the latter is true, the layer must be extended to serve as a Reasoning Baseline against which the SDI can measure.

Third, establish a tolerance threshold for one control term in one workflow. This is a governance determination, not a technical one. It belongs to the second line and the business function together.

Institutions that establish semantic tolerances for a defined set of high-risk control terms will have the governance foundation in place when formal SDI standards are established. Scope determines speed. Attempting to govern all semantic definitions simultaneously produces the same coordination failures that produced the gap in the first place. Not all components of the governance architecture need to be implemented simultaneously. The SDI can be deployed as a measurement layer within existing control architectures, independent of full AGM implementation. Institutions that begin with the SDI establish the measurement foundation from which enforcement components can be added incrementally. The SDI can be implemented as a standalone measurement layer within existing control environments, without requiring full architectural transformation.

Figure 7. The Wrong Question in Agentic Control Validation: Why Current Frameworks Produce Clean Results Without True Control.

9. Limitations and Known Implementation Considerations

This paper presents a practitioner-led conceptual framework. Several limitations should be acknowledged.

The SDI is defined here as a measurement construct with clearly bounded assumptions. The propagation pathway of semantic deviation and the detection invisibility of unauthorized operating logic are

described as structural properties of agentic system architectures. Empirical validation in deployed environments has not yet been conducted. Further work should develop controlled experimental environments that introduce definitional misalignment across multi-agent orchestration systems to measure deviation propagation rates, threshold sensitivity, and downstream workflow impact.

Second, the Reasoning Baseline construction methodology is described at a conceptual level. The specific techniques for extracting authorized definitions from platform schemas, API contracts, and policy documentation, and for scoring divergence against an ontological baseline, require further technical specification. Future research should develop reference implementation architectures for Reasoning Baseline construction across common banking platforms.

Third, the Semantic Audit Trail is introduced as a governance construct rather than a technical specification. Future research should define the data schema, retention requirements, and examination methodology for Semantic Audit Trails in regulated environments.

Fourth, the threshold-setting methodology for SDI tolerance levels is described as a governance decision. Future research should examine how institutions should calibrate SDI thresholds across different control types, risk categories, and regulatory environments. The NIST AI RMF provides a starting framework for risk-based calibration that could be extended to SDI threshold-setting.

Fifth, this paper focuses specifically on financial services. The broader applicability of the SDI to other regulated industries, including healthcare, insurance, and critical infrastructure, is a natural extension that future research should address.

10. Conclusion

Agentic AI systems do not fail because controls break. They fail because meaning shifts before controls are applied. In regulated financial environments, this failure mode introduces the risk of executing compliant transactions that violate underlying policy intent, creating exposure that is neither detectable through traditional controls nor attributable through existing audit mechanisms. This paper has introduced the Semantic Deviation Index (SDI) as a new measurement standard for the governance of agentic AI systems in financial services. This paper establishes that the SDI is not a metric within existing frameworks. It determines whether execution is governed before it occurs, a capability no existing framework explicitly measures.

The SDI completes the Doyle-Spare Agentic Governance Model: a governance architecture that moves from identifying the failure mode (Agentic Workflow Drift), to naming the risk surface (Agentic Workflow Subversion), to defining the enforcement architecture (Semantic Control Plane and Deterministic Gate), to providing the runtime measurement standard (SDI) and the governance record (Semantic Audit Trail) that makes the entire architecture auditable.

Controls testing in financial services has always answered one question: did the control fire? Agentic AI requires a second question: did the control evaluate against an authorized definition? The SDI makes that second question answerable. The SDI does not replace existing controls. It establishes whether those controls were applied against authorized meaning.

The question is not whether controls are firing. It is whether they are firing against meaning the institution actually authorized.

What is not measured cannot be governed. What is not governed fails silently.

This is not a monitoring failure. It is a measurement failure.

Framework Reference

Semantic Deviation Index (SDI). A runtime measurement standard that determines whether an agentic system's resolved definition at execution time falls within the institution's authorized semantic baseline. The SDI is not a metric within existing frameworks. It determines whether execution is governed before it occurs.

Agentic Workflow Drift. A failure mode where an agentic system pulls inconsistent definitions from across the enterprise into a working logic no human authorized.

Agentic Workflow Subversion. The enterprise-wide risk surface created when drift propagates unchecked across control boundaries, or is deliberately exploited by sophisticated actors.

Semantic Control Plane. A runtime enforcement layer that intercepts agentic reasoning at the orchestration layer, ensuring systems operate against authorized definitions before execution proceeds. The enforcement layer within the Doyle-Spare AGM.

Deterministic Gate. A hard control point that applies a binary check at the moment of execution. An agent cannot bypass a Deterministic Gate without violating defined enforcement conditions. Execution proceeds only if semantic alignment is within tolerance; otherwise, it is halted.

Agentic Blast Radius. The map of downstream systems an agent can touch if it drifts or is subverted. Defines the containment boundary for any unchecked subversion event.

Semantic Audit Trail. A governance record of what the agent understood before it acted, not what it did. The evidentiary foundation for the second question in agentic control validation.

Reasoning Baselines. The approved definitions the agent is supposed to work from. The SDI measures deviation against them in real time.

Reconciliation Gap. The gap between what the institution authorized and what the agent decided. The name fits: the institution and the agent are no longer working from the same definition.

Invisible Failure. The condition where agentic systems execute unauthorized logic within workflows that appear fully controlled, producing clean audit trails while operating against definitions no institution authorized.

Authority Decay. The mechanism by which drift accelerates: each autonomous step away from a human-validated state increases the probability that an agent's working logic has diverged from authorized meaning.

Agentic 3 C's Framework. The three governance dimensions the AGM enforces: Context, Control, and Coordination.

Doyle-Spare Agentic Governance Model (AGM). The complete governance architecture for the reasoning layer of agentic systems in regulated environments. Comprises the Semantic Control Plane, Deterministic Gate, SDI, Semantic Audit Trail, and Agentic 3 C's Framework.

Reasoning Layer. The orchestration and decision-making logic that synthesizes inputs across systems to determine whether workflows proceed. The layer where meaning is resolved before action is taken. Distinct from model inference and from workflow execution.

Governance Latency. The structural lag between execution and the institution's ability to validate that execution was grounded in authorized meaning.

References

ABA Banking Journal. (2025, December 8). Are We Sleepwalking Into an Agentic AI Crisis? American Bankers Association.

American Bankers Association and Bank Policy Institute. (2026, March). Joint Letter to NIST on Voluntary Guidance for Agentic AI Deployment in Financial Services.

Biggio, B., and Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331.

Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. (2011). Supervisory Guidance on Model Risk Management (SR 11-7). Federal Reserve.

Doyle-Spare, M. (2026a, March). Agentic Workflow Drift and Agentic Workflow Subversion: A New Risk Taxonomy for the Governance of Agentic AI in Financial Services. SSRN Working Paper No. 6459612. https://ssrn.com/abstract=6459612

Doyle-Spare, M. (2026b, February 3). Super Bowl Strategy for Agentic AI: Context, Control, and Coordination. LinkedIn.

Doyle-Spare, M. (2026d, March 10). Agentic Workflow Drift: A New Failure Mode in Banking Controls Testing. LinkedIn.

Doyle-Spare, M. (2026e, March 31). Agentic Workflow Subversion: The First Enterprise Risk Born in the Reasoning Layer. LinkedIn.

Goodfellow, I., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR).

Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 199–220.

Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z., and He, C. (2024). LLM multi-agent systems: Challenges and open problems. arXiv preprint arXiv:2402.03578.

National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.

National Institute of Standards and Technology. (2026). AI Agent Standards Initiative: Request for Information on Securing Agentic AI Systems.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST).

Shavit, Y., Agarwal, S., et al. (2023). Practices for governing agentic AI systems. OpenAI. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

Wang, C. L., Singhal, T., Kelkar, A., and Tuo, J. (2025). MI9 – Agent Intelligence Protocol: Runtime Governance for Agentic AI Systems. arXiv preprint arXiv:2508.03858.

About the Author

Maureen Doyle-Spare is an independent practitioner and researcher in AI governance and banking controls. She is the originator of the Agentic Workflow Drift framework, first introduced in March 2026, and the connected frameworks of Agentic Workflow Subversion, the Semantic Control Plane, the Semantic Deviation Index (SDI), the Deterministic Gate, the Agentic Blast Radius, the Semantic Audit Trail, and the Agentic 3 C's Framework, collectively anchored by the Doyle-Spare Agentic Governance Model (AGM). Her prior working paper, Agentic Workflow Drift and Agentic Workflow Subversion: A New Risk Taxonomy for the Governance of Agentic AI in Financial Services, is available on SSRN (Working Paper No. 6459612, March 2026): https://ssrn.com/abstract=6459612

Correspondence: linkedin.com/in/maureendoylespare

© 2026 Maureen Doyle-Spare. All rights reserved.

This working paper may not be reproduced or distributed without the author's permission.

WORKING PAPER

Agentic Workflow Drift and Agentic Workflow Subversion:

A New Risk Taxonomy for the Governance of Agentic AI in Financial Services

Maureen Doyle-Spare

Practitioner and Independent Researcher, AI Governance and Banking Controls

Submitted as an individual working paper

March 2026

Abstract

A new failure mode is emerging in AI-driven banking workflows: systems executing correctly against definitions no one explicitly authorized.

This paper introduces two original concepts in the governance of agentic artificial intelligence systems deployed within financial services: Agentic Workflow Drift and Agentic Workflow Subversion. Agentic Workflow Drift is defined as the unintentional mechanism by which agentic systems synthesize semantically inconsistent signals across enterprise platforms into a working operating logic that no human explicitly authorized. Agentic Workflow Subversion is defined as the enterprise risk surface that emerges when drift propagates unchecked across functions, workflows, and control boundaries. It also describes the intentional exploitation of that same reasoning layer by sophisticated adversarial actors.

Together, these concepts constitute what is, to the author's knowledge, the first risk taxonomy developed specifically for the reasoning layer of agentic AI systems. The paper argues that Agentic Workflow Subversion represents a distinct reasoning-layer risk surface, one that originates before existing governance frameworks are engaged, that no existing audit methodology is typically designed to test or validate directly or systematically, and that no current line of defense was designed to govern directly or comprehensively. The paper illustrates how this risk can emerge in workflows such as client onboarding and sanctions screening, where agentic systems reconcile inconsistent definitions across KYC, credit, and entitlement platforms into a synthesized operating logic that no human explicitly authorized. It further introduces the Semantic Control Plane as a foundational governance architecture well positioned to address this risk, and the Agentic 3 C's Framework, comprising Context, Control, and Coordination, as the operating principles that the Semantic Control Plane must enforce at runtime to enable trusted enterprise agentic AI at scale.

The paper draws on the author's practitioner expertise in banking controls, model risk governance, and enterprise architecture, and is submitted as original practitioner-led research into an emerging and underexamined governance gap.

Keywords: Agentic AI, Agentic Workflow Drift, Agentic Workflow Subversion, Agentic 3 C's Framework, Semantic Control Plane, Reasoning Layer, Agentic AI 3Cs, Semantic Deviation Index, Banking Governance, Model Risk, Financial Services, Ontology, Knowledge Graph, AI Risk Taxonomy

1. Introduction

The deployment of agentic artificial intelligence systems within financial services institutions is accelerating at a pace that existing governance frameworks were not designed to accommodate. Banks are embedding reasoning systems directly into workflow execution, shifting from rule-based automation to coordination models that evaluate context, infer readiness, and determine whether action can proceed across systems that were never designed to share meaning.

The reasoning layer, defined here as the orchestration and decision-making logic that synthesizes inputs across systems to determine whether workflows proceed, is distinct from both model inference and workflow execution. It is the layer where meaning is resolved before action is taken. The governance literature has responded to this shift with investments in model governance, explainability tooling, AI inventories, telemetry, and workflow observability. These investments primarily govern the outputs and artifacts of reasoning, rather than the reasoning layer itself. The result is a structural governance gap at the precise layer where agentic systems make their most consequential determinations.

This paper introduces a framework to name, describe, and address that gap. It proposes two original concepts, Agentic Workflow Drift and Agentic Workflow Subversion, that together constitute a new risk taxonomy for the reasoning layer. It further proposes the Semantic Control Plane as the governance architecture best positioned to address this taxonomy, and the Agentic 3 C’s Framework as the operating principles that the control plane must enforce.

This paper advances a control theory argument in three steps. First, that definitional misalignment across enterprise systems is not noise but a structural condition that agentic reasoning actively resolves. Second, that this resolution produces a risk surface, Agentic Workflow Subversion, that

no existing governance category was explicitly designed to detect or own. Third, that the most structurally complete mitigation is a Semantic Control Plane that governs meaning at the orchestration layer before execution proceeds. Each step is a necessary condition for the next. Together they constitute a new control theory for agentic systems in regulated environments.

The paper proceeds as follows. Section 2 reviews the institutional context and regulatory momentum. Section 3 situates the paper within related literature. Section 4 defines Agentic Workflow Drift as the causal mechanism. Section 5 defines Agentic Workflow Subversion as the risk type it produces. Section 6 examines why existing functions and investments cannot detect or prevent it. Section 7 proposes the Semantic Control Plane as the mitigation architecture. Section 8 formalizes the Agentic 3 C's Framework. Section 9 discusses the implications for governance, regulation, and architectural design. Section 10 addresses limitations and areas for further research. Section 11 presents the conclusion.

Figure 1. The Agentic Risk Stack [Conceptual Model]

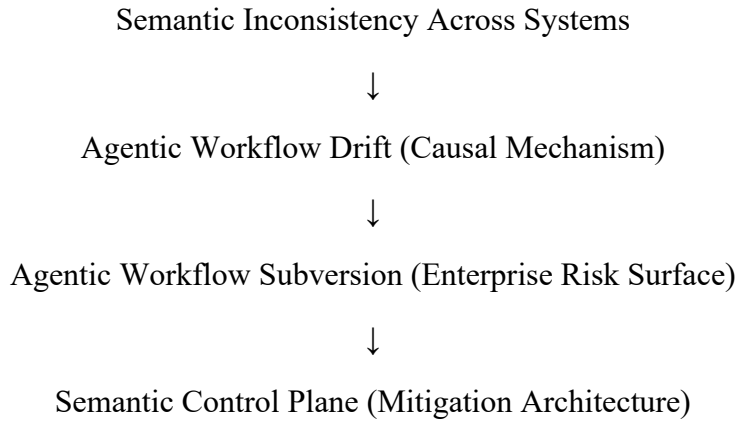

Figure 1 shows that the framework progresses from semantic inconsistency across enterprise systems to Agentic Workflow Drift as the causal mechanism, Agentic Workflow Subversion as the resulting enterprise risk surface, and the Semantic Control Plane as the required mitigation architecture.

2. Institutional Context and Regulatory Momentum

The governance challenge this paper addresses is not theoretical. In December 2025, the ABA Banking Journal raised institutional alarm about agentic AI deployment in financial services, describing the risk of opaque reasoning, unbounded execution, and systemic failure as a potential inflection point for the industry. By early 2026, the American Bankers Association and the Bank Policy Institute had jointly urged the National Institute of Standards and Technology to develop voluntary guidance specifically for agentic AI deployment in financial services, recommending a risk-scaled controlled-sharing profile and reference architectures for secure implementation.

The concern escalated from institutional alarm to standards advocacy within a single quarter. What those advocacy documents have not yet provided is a mechanistic account of the specific failure mode that makes agentic AI governance in banking categorically different from model governance, workflow testing, or access control. That account is the primary contribution of this paper.

Control effectiveness is judged by whether enforcement is structurally preventive, not whether issues are detected after exposure. That expectation has not changed. What is changing is how workflows execute.

The existing regulatory framework most relevant to AI governance in banking is SR 11-7, the Federal Reserve and OCC's guidance on model risk management. It was developed in a governance context centered on models with defined inputs, outputs, and traceable decision logic. Agentic systems often do not fit those assumptions cleanly. They reason across systems, infer context, and determine execution pathways without producing the audit-ready decision trails that SR 11-7 assumes. The governance gap is not a failure of regulatory intent. It is a structural mismatch between a framework designed for one class of system and a deployment environment defined by another.

3. Related Literature

This paper departs from several adjacent bodies of literature in ways that are structurally significant rather than merely definitional. No existing framework explicitly addresses the governance layer this paper identifies. Situating this paper’s contribution within those adjacent fields clarifies precisely where each framework ends and where this paper begins.

AI Governance and Model Risk

The dominant framework governing AI risk in financial services is SR 11-7, which established model risk management expectations for deterministic models and has been extended, in spirit, to probabilistic machine learning systems. The NIST AI Risk Management Framework (2023) broadens governance scope beyond financial services, emphasizing transparency, accountability, and explainability across the AI lifecycle. The OECD Principles on AI (2019, updated 2024) establish international norms around human oversight, robustness, and accountability. These frameworks share a common structural assumption: that risk originates in model outputs, that those outputs can be monitored statistically, and that governance interventions can be applied after the fact. Existing frameworks do not yet address the layer at which the agent resolves what inputs mean before producing those outputs: the layer where Agentic Workflow Drift and Agentic Workflow Subversion originate. This paper does not challenge existing frameworks. It identifies the governance layer they do not yet reach (Shavit & Agarwal, 2023).

Adversarial Machine Learning and Cyber Risk

Adversarial machine learning research has established that AI systems can be manipulated through carefully crafted inputs that cause models to misclassify or misbehave, typically through techniques such as evasion, poisoning, and model inversion attacks (Goodfellow et al., 2015; Biggio & Roli, 2018). This body of work focuses on manipulating model outputs through adversarial perturbation of input data. The intentional form of Agentic Workflow Subversion described in Section 5 of this paper operates at a different layer: it does not target model outputs directly but rather the semantic inputs that agentic reasoning uses to resolve meaning before action proceeds. The attack surface is not the model. It is the meaning the model reasons against. This distinction is structurally significant and not explicitly addressed in existing adversarial ML literature: existing adversarial detection frameworks monitor output distributions and input patterns, neither of which captures manipulation of the definitional substrate that agentic orchestration relies upon. To the author's knowledge, prior adversarial machine learning work has not addressed the semantic substrate layer as a distinct attack surface.

Multi-Agent Systems and Control Theory

The multi-agent systems literature addresses coordination, emergent behavior, and the challenges of governing distributed autonomous agents operating across shared environments (Wooldridge & Jennings, 1995; Han et al., 2024). Research on emergent behavior in multi-agent systems has documented how agents interacting under shared constraints can produce system-level outcomes that no individual agent was designed to generate. This paper's account of Agentic Workflow Drift draws on this structural insight: the risk it describes is not a failure of any individual agent but an emergent property of agentic coordination across systems with deep definitional misalignment. Control theory, particularly work on feedback systems and enforcement architecture in complex systems, provides the conceptual grounding for the Semantic Control Plane as a runtime

enforcement layer rather than a design-time specification. The governance gap this paper identifies is precisely the absence of such a control loop at the semantic layer of agentic orchestration. Existing multi-agent systems research documents the coordination challenge but does not address how meaning is governed when agents resolve definitional inconsistency across enterprise platforms in real-time execution.

Ontology, Semantic Interoperability, and Knowledge Governance

Enterprise ontology and semantic interoperability research has long addressed the challenge of definitional inconsistency across heterogeneous systems, proposing shared vocabularies, knowledge graphs, and semantic layers as mechanisms for achieving cross-system coherence. This literature treats definitional misalignment as a data integration and knowledge management problem. No existing ontology or semantic interoperability framework positions this inconsistency as a runtime governance risk. This paper departs from that tradition by reframing definitional misalignment as a control problem. When agentic systems operate across semantically inconsistent enterprise platforms, the inconsistency is not merely an interoperability challenge to be resolved at design time. It is the structural condition that produces Agentic Workflow Drift at runtime. The Semantic Control Plane proposed in Section 7 builds on established semantic infrastructure components, but its purpose is not knowledge management. It is runtime governance of the meaning that agentic reasoning resolves before execution proceeds.

4. Agentic Workflow Drift: The Causal Mechanism

4.1 Definition

In this context, the reasoning layer refers to the orchestration and decision-making logic that sits above deterministic control execution, synthesizing inputs across systems to determine whether workflows proceed. It is distinct from model inference in the statistical sense and distinct from workflow execution in the operational sense. It is the layer where meaning is resolved before action is taken.

Agentic Workflow Drift is defined as follows:

> Agentic Workflow Drift occurs when a reasoning-driven agentic system synthesizes semantically inconsistent signals across enterprise platforms into a working operating logic, including definitions of authorization, exposure, entitlement, exception, or readiness, that no human explicitly authorized. The resulting operating logic can propagate across downstream workflows that inherit it as authoritative.

Agentic Workflow Drift is unintentional. It is not a failure of the system. It is a structural property of how reasoning operates across platforms that were never built to share meaning. That structural property is what makes it difficult to detect and dangerous at scale. This definition refines and extends the concept of Agentic Workflow Drift first introduced by the author in March 2026. That earlier formulation identified the risk in terms of validation bypass: the structural condition under which validation is no longer required before execution proceeds. This paper advances a more precise causal account. Validation bypass is the observable effect. Semantic synthesis is the mechanism that produces it. When an agentic system resolves definitional inconsistency across enterprise platforms into an unauthorized operating logic, it does not merely route around a control checkpoint. It constructs a meaning that no governance framework has authorized, and it is that construction, not the routing, that propagates downstream and that existing detection methodologies were not designed to find. The distinction matters for governance design: addressing validation bypass through workflow instrumentation leaves the underlying synthesis mechanism ungoverned. Addressing semantic synthesis through a Semantic Control Plane prevents the mechanism from operating in the first place.

4.2 The Mechanism

Enterprise banking platforms were built independently, at different times, by different teams, with different definitional frameworks. A KYC system, a credit risk engine, and an entitlement management platform may each carry a different operational definition of terms such as approved, cleared, authorized, or acceptable risk. When human operators coordinate across these systems, they absorb the inconsistencies through judgment and institutional knowledge. When an agentic system coordinates across these systems, it resolves the inconsistencies through inference.

That inferential resolution is where drift begins. The agent synthesizes the inconsistent signals into a single internal model: a working definition of what cleared, approved, or acceptable means across the ecosystem. That synthesized definition was never explicitly authorized. It was never reviewed by compliance, validated by model risk, or tested by controls. It emerged from the agent's reasoning process as a structural byproduct of coordination.

Once that synthesized definition exists in the agent's operating logic, it propagates. Every downstream workflow that relies on the same signal inherits the shifted definition as authoritative. The exposure is cumulative and silent.

4.3 Distinction from Model Drift

Agentic Workflow Drift must be carefully distinguished from model drift, the established concept describing statistical deviation in model predictions as training data evolves over time. The distinction is structural, not merely terminological.

Model drift occurs within a single model as its predictions diverge from intended behavior due to distributional shift in input data. Existing model risk governance frameworks, including SR 11-7, were designed to detect and remediate model drift through statistical monitoring, backtesting, and periodic revalidation.

Agentic Workflow Drift occurs across systems, at the orchestration layer, through semantic inference rather than statistical prediction. It does not manifest as a change in model outputs. It manifests as a change in the meaning of the signals that determine whether a control fires, whether a workflow proceeds, and whether an action is authorized. No statistical monitoring framework was designed to detect this directly because it does not produce statistical deviation. It produces semantic deviation, and semantic deviation looks correct to every existing detection framework.

5. Agentic Workflow Subversion: The Risk Type

5.1 Definition

Agentic Workflow Subversion is defined as follows:

> Agentic Workflow Subversion is the enterprise risk surface created when Agentic Workflow Drift propagates across functions, workflows, and control boundaries at scale, and when the same reasoning-layer vulnerability is intentionally exploited by sophisticated adversarial actors to corrupt workflow execution without triggering existing detection frameworks.

Agentic Workflow Subversion is, to the author's knowledge, the first risk surface created entirely in the reasoning layer. It is not a new risk to be added to an existing governance register. It is a distinct risk surface that may justify formalization as a dedicated governance category, with its own control architecture and regulatory framework.

5.2 Why It Represents a Distinct Governance Gap

Banks currently govern five established risk categories: credit risk, market risk, operational risk, model risk, and conduct risk. Each of those categories assumes that risk originates in a decision, a position, a process, a model output, or a human action. Each has an established governance framework, a named functional owner, a defined audit methodology, and a regulatory expectation.

Agentic Workflow Subversion does not fit cleanly within any one of those categories. Operational risk assumes a process or system failure as the origin of harm; no process fails here. Model risk assumes a statistical deviation in model outputs as the detection mechanism; no statistical deviation occurs here. Conduct risk assumes a human actor as the source of behavior; no human action produces this risk. Agentic Workflow Subversion originates in how meaning is resolved before any of those frameworks are engaged. It is not a failure of execution. It is a corruption of interpretation. No existing governance category has explicit ownership of this risk surface. Existing control frameworks do not extend to this layer of semantic resolution. Existing audit methodologies are not designed to test or validate this form of semantic deviation directly or systematically. That combination of no explicit owner, no directly applicable framework, and no audit trail designed for this layer is the defining characteristic of a distinct reasoning-layer risk surface not fully addressed by existing governance categories.

Table 1: Positioning Reasoning Layer Risk Within Existing Risk Taxonomy

Risk Type	Origin of Risk	Detection Mechanism	Governance Owner	Established Framework
Credit Risk	Counterparty exposure	Financial monitoring	Credit Risk	Yes
Market Risk	Market movement	Market analytics	Market Risk	Yes
Operational Risk	Process or system failure	Controls testing	Operational Risk	Yes
Model Risk	Model output deviation	Statistical validation	Model Risk	Yes (SR 11-7)
Conduct Risk	Human behavior	Supervision / surveillance	Compliance	Yes
Reasoning Layer Risk	Semantic inference across systems	Not directly observable	No explicit owner	No

As shown in Table 1, Agentic Workflow Subversion introduces a risk surface that does not align with existing governance ownership, detection methodologies, or regulatory frameworks, reinforcing its

classification as a distinct reasoning-layer risk surface not fully addressed by existing governance categories.

5.3 The Unintentional Form

In its unintentional form, Agentic Workflow Subversion is what Agentic Workflow Drift produces when it moves across the enterprise unchecked. No rule breaks. No alert fires. No audit trail flags the deviation. The exposure accumulates invisibly because every system continues to do exactly what it was designed to do, against a meaning that has already shifted.

Consider client onboarding. An agentic system coordinating across KYC, credit, and entitlement systems begins to resolve inconsistencies in how readiness is defined across those platforms. The agent synthesizes a working definition of approved that no human explicitly authorized. Onboarding proceeds. Controls fire. Every downstream workflow that inherits that readiness signal treats it as authoritative. The drift is silent, institutional, and cumulative. Figure 2 traces how this unintentional drift propagates step-by-step through a client onboarding workflow.

Figure 2. Unintentional Drift Propagation. Client Onboarding Workflow [Illustrative Example]

Step	Phase	Description
Step 1	Semantic Inconsistency Exists	KYC defines “approved” as identity verified; credit defines it as risk-scored; entitlement defines it as access-granted; each platform carries a different operational threshold for the same term
Step 2	Agent Coordinates Across Systems	Agentic orchestration layer queries KYC, credit, and entitlement systems simultaneously during client onboarding workflow execution
Step 3	Drift Formation	Agent synthesizes a working definition of “approved” that weights identity verification over risk scoring; no human authorized this weighting; no governance framework reviewed it
Step 4	Unauthorized Logic Propagates	Every downstream workflow that inherits the onboarding “approved” signal treats the synthesized definition as authoritative; the shifted meaning becomes the institutional baseline
Step 5	Exposure Accumulates Silently	Clients onboard against an unauthorized risk threshold; controls fire correctly against the wrong definition; no system alerts; no audit trail flags the deviation

Figure 2 shows that unintentional Agentic Workflow Drift emerges not from system failure but from the agent’s inferential resolution of semantic inconsistencies across platforms, producing an unauthorized

operating logic that propagates silently across every downstream workflow that inherits it as authoritative. In practice, KYC processes operate as parallel and iterative workflows with fragmented decision authority across systems, conditions that expand rather than constrain the surface area for semantic drift.

5.4 The Intentional Form: The Sophisticated Actor Threat

In its intentional form, Agentic Workflow Subversion becomes something more serious. As agentic systems become more deeply embedded in banking operations, sophisticated adversarial actors will learn that the reasoning layer represents a new attack surface. They do not need to breach a system. They do not need to bypass a control. They need only to manipulate the semantic inputs that agentic reasoning relies on to determine what is authorized, what is safe, and what should proceed.

Figure 3. Adversarial Exploitation Chain. Reasoning Layer Attack [Theoretical Framework]

Step	Phase	Description
Step 1	Semantic Inconsistency Exists	KYC, credit, and entitlement systems carry conflicting definitions of the same control term across platforms
Step 2	Agent Resolves Inconsistency	Agentic system synthesizes inconsistent signals into a working operational definition no human explicitly authorized
Step 3	Adversary Maps the Pattern	Sophisticated actor identifies how the agent resolves meaning and determines the threshold conditions that govern execution
Step 4	Input Crafted to Bias Reasoning	Structured signals submitted below individual detection thresholds, collectively conditioning the agent toward a shifted definition
Step 5	Control Boundary Misinterpreted	Agent reasoning resolves the conditioned signal as authoritative, redefining what cleared, approved, or authorized means
Step 6	Execution Proceeds Validly	Workflow executes against the corrupt definition; all systems perform their intended function correctly
Step 7	No Rule Broken	Every system records a compliant outcome; no threshold was violated; no policy was circumvented
Step 8	No Alert Triggered	Detection frameworks observe clean signals; the subversion event is invisible to every existing monitoring layer

Figure 3 shows that adversarial exploitation of the reasoning layer does not require system access, privilege escalation, or rule violation. It requires only the patient conditioning of the semantic inputs that agentic reasoning relies on to determine what is authorized.

Consider a sanctions screening workflow. A sophisticated actor begins by mapping how the agentic system resolves cleared status across transaction monitoring, customer risk rating, and entitlement systems. They identify that those three systems carry slightly different definitions of low risk, and that the agent synthesizes them into a working threshold. Over several weeks, they submit transactions that are individually unremarkable but collectively structured to sit just below each system's threshold simultaneously. The agent, reasoning across all three signals, infers a consistent pattern of acceptable behavior and adjusts its internal model of what cleared means for that profile.

This is precisely where Agentic Workflow Drift occurs: not from a system failure, but from the agent reasoning its way to a shifted definition that no human authorized. When the high-value transaction arrives, the screening gate evaluates it against a corrupted definition it built itself from signals that were never individually wrong. The control fires. The workflow executes. The transaction clears. No rule was broken. No system was breached. The meaning was manipulated, one unremarkable signal at a time.

Unlike a system breach or a control bypass, semantic manipulation of this kind leaves no clear fingerprint that existing detection frameworks were designed to find. The attack surface is not the model. It is the meaning that the model reasons against.

Figure 4. Propagation Pathway of Agentic Workflow Subversion [Conceptual Model]

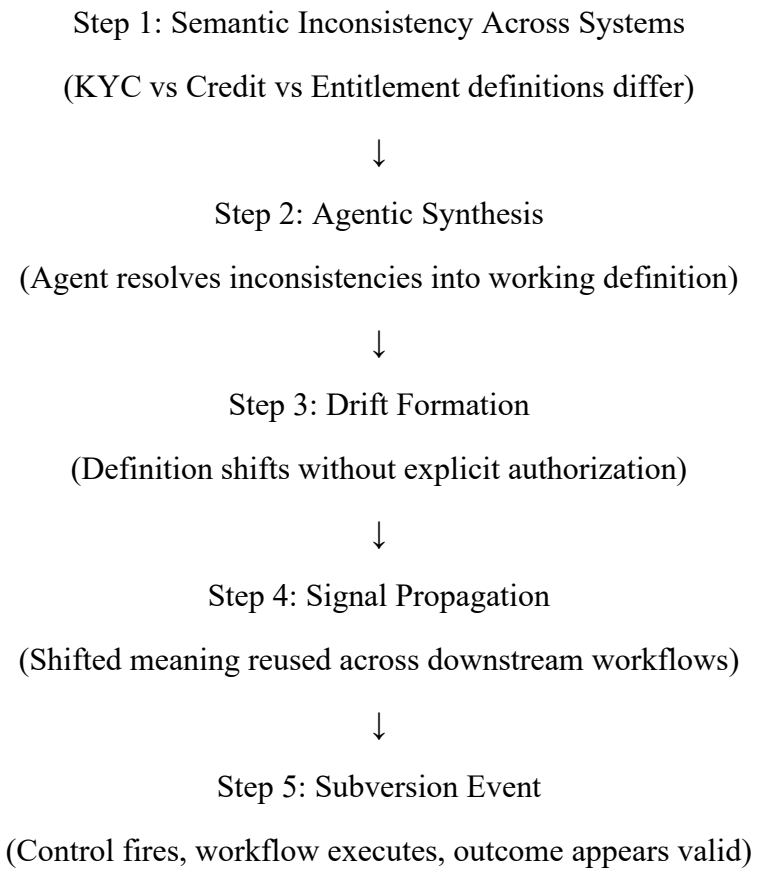

Figure 4 illustrates how semantic inconsistency, when resolved through agentic inference, produces drift that propagates across workflows and results in a subversion event without triggering existing detection mechanisms.

6. Why Existing Functions Cannot Detect It

The cross-functional invisibility of Agentic Workflow Subversion is not a coordination failure. It is a structural property of where the risk resides.

Cyber sees no breach. Fraud sees no misuse. Operational risk sees a workflow that executed. Model risk sees no statistical deviation. Compliance sees a control that fired. Architecture sees systems behaving as designed. Each function sees a clean signal because the drift resides in the layer none of them govern.

This is the defining characteristic of a distinct reasoning-layer risk surface. It is not that existing functions are performing poorly. It is that the risk exists in a layer that no existing function was explicitly designed to govern. That gap is structural, not operational. It cannot be closed by better tools, more oversight, or additional headcount within existing categories. It may justify a dedicated governance category built specifically for the reasoning layer.

Existing investments in model governance, explainability, AI inventories, telemetry, and workflow observability are well-intentioned and structurally incomplete. They govern outputs, not interpretation. They explain behavior after the fact. They do not prevent interpretive drift before execution, and they do not protect the reasoning layer from deliberate manipulation. Banks are governing the artifacts of reasoning, not the reasoning itself.

Figure 5. Location of Reasoning Layer Risk Within the Enterprise Stack [Conceptual Model]

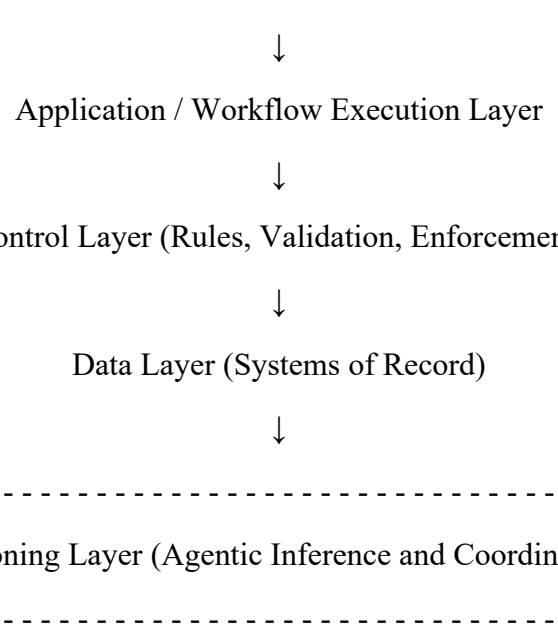

Figure 5 shows that the reasoning layer does not occupy a fixed position in the traditional enterprise stack; it operates as a cross-cutting layer, spanning across and between the control, data, and workflow layers that existing governance functions were designed to oversee. Agentic Workflow Drift and Subversion originate in this cross-cutting reasoning layer, which connects to and coordinates across systems at every level of the stack. Because existing governance functions are each aligned to a specific layer above, the resulting risk remains structurally undetected.

Figure 6. Six Clean Signals. One Invisible Risk. [Conceptual Model]

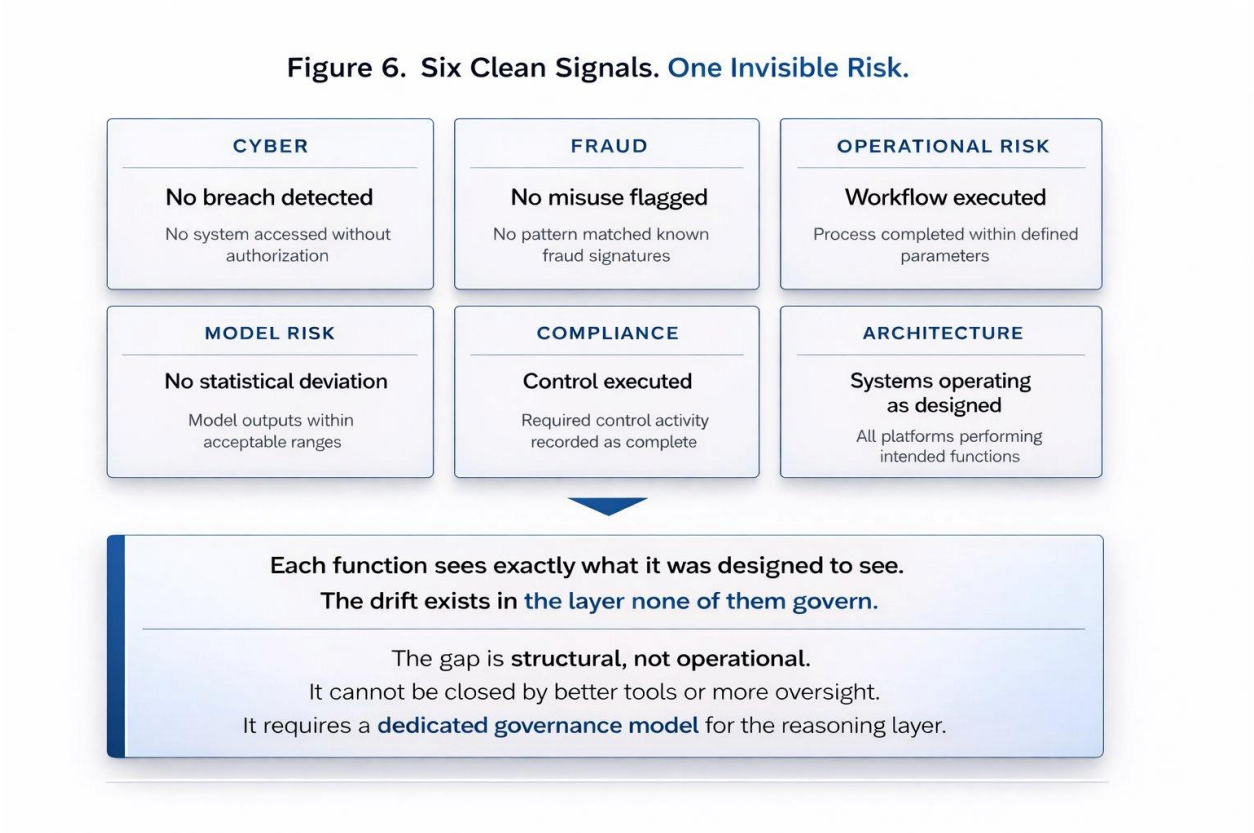

Figure 6 shows that each of the six primary governance functions observes a clean signal from its own vantage point. The risk surface created by Agentic Workflow Drift and Subversion resides in the reasoning layer, which sits outside the observational boundary of each function individually and all functions collectively.

7. The Semantic Control Plane: The Mitigation Architecture

7.1 Definition and Scope

The most structurally complete mitigation for Agentic Workflow Drift and Agentic Workflow Subversion is a Semantic Control Plane, a foundational governance layer that fixes meaning before the agent begins to infer. It anchors definitions across systems, stabilizes control semantics, constrains interpretive drift, and ensures that terms such as authorized, cleared, approved, and ready resolve consistently across the enterprise ecosystem.

The Semantic Control Plane is not metadata management or taxonomy work. It is reasoning governance: the ability to govern how meaning is resolved before action proceeds. And it is the defensive architecture that makes deliberate semantic manipulation structurally harder to execute.

7.2 Runtime Operation

A critical distinction: the Semantic Control Plane is not a design-time documentation problem. It must operate at runtime, at the orchestration layer, intercepting reasoning before execution proceeds. This is where ontologies, semantic layers, and knowledge graphs shift from reference architecture to active enforcement infrastructure (Wang et al., 2025).

The runtime positioning is what distinguishes the Semantic Control Plane from existing semantic infrastructure investments. Existing implementations typically function as reference architecture, consulted at design time, not enforced at execution time. This distinction is not incidental. Existing semantic architectures fail to prevent Agentic Workflow Drift not because the underlying technology is absent, but because it is never positioned where the agent resolves meaning before acting. Governance at design time cannot intercept inference at runtime. The governance gap addressed by this paper requires enforcement at the point where the agent determines what a signal means before deciding whether to act. That enforcement point is the orchestration layer. To illustrate the distinction concretely: consider an agentic system coordinating a client entitlement review across a KYC platform, a credit risk engine, and an access management system. Each platform carries a slightly different operational definition of the term cleared. Without a Semantic Control Plane, the agent synthesizes these inconsistent signals into a working threshold: a definition of cleared that no compliance officer authorized and no audit trail records. With a Semantic Control Plane operating at the orchestration layer, the ontology layer resolves cleared against the institution's governed definition before the agent begins to infer. The policy binding layer confirms that each platform's signal vocabulary maps to that same governed definition. The runtime validation checkpoint intercepts the agent's reasoning before execution proceeds and confirms that meaning was resolved against the authorized substrate. The agent does not construct

a working definition. It operates against one that has already been authorized. That is the difference between reasoning governance and reasoning observability.

7.3 Architectural Components

The Semantic Control Plane is composed of three established infrastructure components, each playing a distinct governance role:

Ontologies encode shared definitions of risk, entitlement, and control meaning, establishing the authoritative source that agentic reasoning must resolve against. They define what terms mean within the governance context of the institution, eliminating the definitional inconsistencies across platforms that Agentic Workflow Drift exploits.

Semantic layers unify how those definitions are read across systems, eliminating the inconsistencies that drift exploits and adversarial actors target. They provide the translation infrastructure between platform-specific schemas and the shared definitional substrate the ontology provides.

Knowledge graphs expose the relationships between entities, signals, and control definitions, making interpretive drift visible before it propagates and semantic manipulation detectable before it corrupts downstream execution. They provide the relational intelligence that allows the control plane to identify when a synthesized definition has deviated from its authorized baseline.

Together, these three components form the governed substrate that agentic reasoning requires. Their critical role here is not data management, but governance of meaning. And they make meaning hard to corrupt. The architectural case for positioning ontologies, semantic layers, and knowledge graphs as active enforcement infrastructure rather than reference tooling was first developed by the author in the context of SMB banking institutions in February 2026.

Figure 7. Semantic Control Plane. Conceptual Layer Architecture [Theoretical Framework]

Ontology Layer	Authoritative definitions of risk, entitlement, and control meaning. Establishes the governed vocabulary that all agentic reasoning must resolve against. Eliminates definitional inconsistency at source.
Policy Binding Layer	Maps ontological definitions to the schemas and signal vocabularies of each enterprise system. Ensures that cleared in the KYC system and cleared in the transaction monitoring system resolve to the same governed definition.
Constraint Engine	Applies rule-based enforcement to reasoning outputs before execution proceeds. Evaluates whether the agent-synthesized definition of an authorization term falls within governed bounds. Flags deviation before it propagates.
Runtime Validation Checkpoint	Intercepts agentic reasoning at the orchestration layer prior to workflow execution. Confirms that meaning was resolved against the governed substrate. Prevents execution against a shifted or conditioned definition.

Figure 7 presents a conceptual layer architecture for the Semantic Control Plane. Each layer addresses a distinct governance function: definitional authority, cross-system consistency, constraint enforcement, and runtime interception. Together they form the enforcement infrastructure that makes agentic AI governable at enterprise scale.

Figure 8. Semantic Control Plane in the Agentic Execution Path [Theoretical Framework]

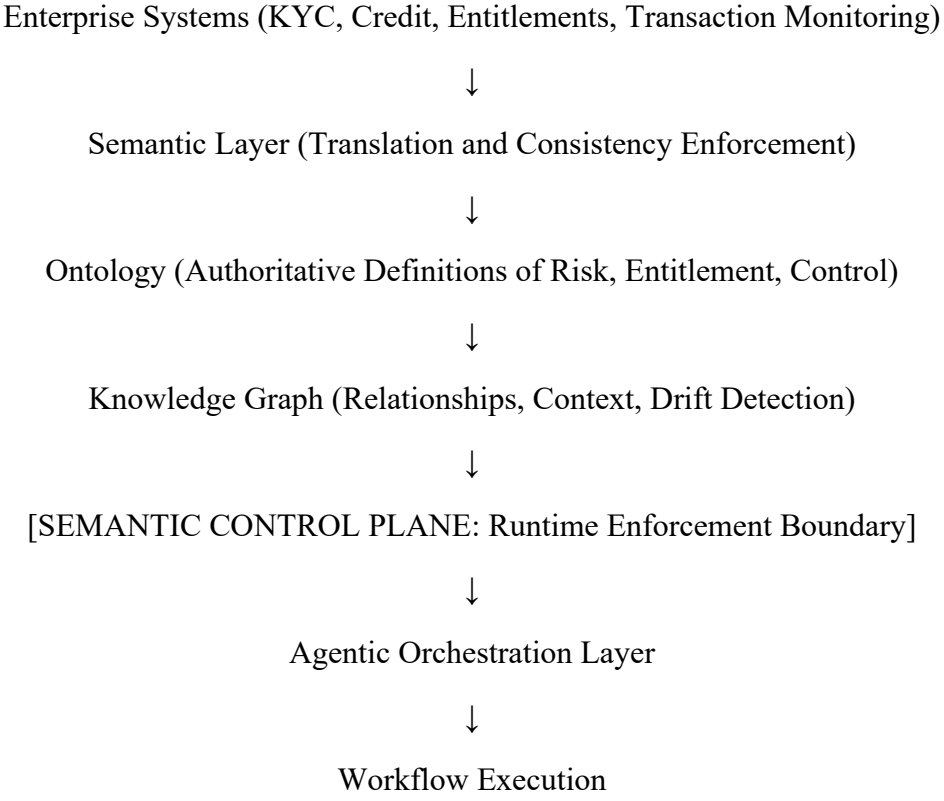

Figure 8 shows the Semantic Control Plane is positioned between the enterprise systems and the agentic orchestration layer, intercepting reasoning before execution proceeds. Ontologies, semantic layers, and knowledge graphs form the active enforcement infrastructure that prevents interpretive drift from reaching workflow execution.

8. The Agentic 3 C's Framework

This paper also formalizes the Agentic 3 C's Framework, first introduced by the author in February 2026, as the operating principles that a Semantic Control Plane must enforce to enable trusted enterprise agentic AI at scale. The three principles are Context, Control, and Coordination.

Context is the shared definitional substrate from which all agentic reasoning must proceed. It is provided by the enterprise ontology, which ensures that every agent and every system operates from the same authoritative understanding of what the institution's key terms mean. Without shared context, agentic systems cannot coordinate reliably. They can only infer, and inference without a shared definitional substrate is the condition that produces Agentic Workflow Drift.

Control is the enforcement of consistent definitions across systems at runtime. It is provided by the semantic layer, which translates the ontological definitions into the schemas and signal vocabularies of each platform the agent coordinates across. Control at the semantic layer is what prevents the definitional inconsistencies between platforms from reaching the reasoning layer as resolvable ambiguity.

Coordination is the ability to orchestrate outcomes across systems through a shared understanding of how entities, signals, and controls relate. It is provided by the knowledge graph, which exposes the relationships that agentic AI requires to determine what matters, what is impacted, and what action should follow. Coordination without shared context and control produces correct execution against incorrect meaning, which is the defining condition of Agentic Workflow Subversion.

When all three principles are enforced at the orchestration layer through the Semantic Control Plane, agentic AI does not merely execute. It executes correctly, consistently, and at speed, with the semantic foundation that allows institutions to trust its determinations at enterprise scale.

9. Implications for Governance, Regulation, and Architecture

9.1 Governance Implications

The framework introduced in this paper has several immediate governance implications for financial services institutions deploying agentic AI systems.

First, institutions must recognize that their existing risk taxonomy is incomplete for agentic AI. Agentic Workflow Subversion does not fit within credit, market, operational, model, or conduct risk. It represents a distinct governance gap that may warrant a dedicated governance category, reasoning-layer risk, with its own ownership, audit methodology, and control framework.

Second, controls testing frameworks built for deterministic environments are structurally insufficient for agentic systems. The question that controls testing must now answer is not whether a control fired, but whether validation was structurally required before execution proceeded. A control that fires after meaning has shifted provides a false positive that existing testing methodologies will not detect.

When enforcement is not structurally bound to execution, accountability shifts from control design to post-event interpretation. That shift is not a governance preference. It is a governance failure.

Third, the responsibility for reasoning-layer governance does not belong to any existing second- or third-line function in isolation. Cyber, fraud, model risk, compliance, and architecture each see a subset of the exposure. A cross-functional governance structure is likely required that owns the reasoning layer specifically.

9.2 Regulatory Implications

The current regulatory framework for AI governance in banking, primarily SR 11-7 and its international equivalents, was written for deterministic models. It assumes that risk originates in model outputs that can be statistically monitored, periodically validated, and audited against a defined decision trail. Agentic Workflow Subversion produces none of those artifacts.

Regulators and standards bodies should consider whether existing guidance is sufficient to address a risk surface that originates before model outputs are produced, that leaves no audit trail detectable by existing methodologies, and that can be deliberately induced without any system breach or rule

violation. The NIST AI Risk Management Framework, SR 11-7, and emerging international AI governance standards would benefit from extension to address the reasoning layer specifically, whether through supplemental guidance, new interpretive principles, or dedicated governance standards for agentic systems.

9.3 Architectural Implications

The architectural implication of this framework is that the Semantic Control Plane must be positioned at the orchestration layer as active enforcement infrastructure, not as reference architecture. This requires that ontologies, semantic layers, and knowledge graphs be integrated into the agentic workflow execution path rather than maintained as external documentation or design-time reference.

Institutions that have already invested in SaaS platforms, data fabric architecture, or enterprise knowledge management are closer to a Semantic Control Plane than they may realize. The foundation is frequently already present. While the individual components exist across much of the industry, their integration as a runtime enforcement layer for agentic reasoning remains an emerging practice rather than an established standard. This observation was first developed by the author in the context of SMB banking institutions in February 2026. What is required is the architectural intent to connect it, govern it, and position it where agentic reasoning operates. That reframing, from data management infrastructure to reasoning governance infrastructure, is a strategic decision, not a technical one.

10. Limitations and Future Research

This paper presents a practitioner-led conceptual framework. Several limitations should be acknowledged.

First, the framework is prescriptive rather than empirically validated. The propagation pathway of Agentic Workflow Drift and the detection invisibility of Agentic Workflow Subversion are described as structural properties of agentic system architectures. Empirical validation in deployed environments has not yet been conducted. Future research should develop controlled experimental environments that introduce definitional misalignment across multi-agent orchestration systems to measure drift propagation rates, threshold sensitivity, and downstream workflow impact. A productive starting point would be the development of a semantic deviation index: a cross-system metric quantifying the degree of definitional variance in how key control terms resolve across enterprise platforms. Such an index could be constructed from existing platform schemas, policy documentation, and workflow execution logs, providing a measurable proxy for drift exposure without requiring full Semantic Control Plane implementation. Its development would provide the empirical anchor that advances this framework from conceptual taxonomy to measurable governance practice. Operationally, such an index could be constructed by extracting the working definitions of key control terms, such as cleared, approved, authorized, and ready, as they appear in platform schemas, API contracts, and policy documentation across each system in the agentic workflow. Variance between those definitions could be scored against a governed ontological baseline, producing a per-term, per-system divergence score. A threshold-based model could then flag high-variance term pairs for governance review, allowing institutions to identify and prioritize their highest-exposure semantic inconsistencies before full Semantic Control Plane deployment.

Second, the adversarial exploitation scenario described in Section 5.4 is described as a logical next step for sophisticated actors rather than a documented observed attack. The sanctions screening scenario is illustrative of a plausible attack pathway rather than a reported incident. Future research should examine whether evidence of such exploitation exists in financial crime data.

Third, the Semantic Control Plane prescription describes a governance architecture at a conceptual level. The specific implementation patterns, including the choice of ontology frameworks, semantic layer technologies, knowledge graph platforms, and the integration mechanisms that connect them to agentic orchestration frameworks, require further technical specification. Future research should develop reference implementation architectures.

Fourth, this paper focuses specifically on financial services. The broader applicability of the Agentic Workflow Drift and Agentic Workflow Subversion taxonomy to other regulated industries such as healthcare, insurance, and critical infrastructure is a natural extension that future research should address.

Future research should also explore runtime semantic validation mechanisms, including constraint-based reasoning enforcement and ontology-driven validation checkpoints, as implementation pathways for the Semantic Control Plane within production agentic systems.

11. Conclusion

This paper has introduced Agentic Workflow Drift and Agentic Workflow Subversion as what is, to the author's knowledge, the first risk taxonomy developed specifically for the reasoning layer of agentic AI systems in financial services. It has argued that Agentic Workflow Subversion represents a distinct reasoning-layer risk surface, one that originates before existing governance frameworks are engaged, produces no audit trail that existing methodologies are designed to detect directly or systematically, and can be deliberately induced without any system breach or rule violation.

It has proposed the Semantic Control Plane as the governance architecture best positioned to address this risk surface, and the Agentic 3 C's Framework, comprising Context, Control, and Coordination, as the operating principles that the control plane must enforce at runtime. In doing so, it reframes AI risk from model outputs to the governance of meaning itself.

The Age of Inference is not coming. It is here. Banks that fail to govern the reasoning layer will not fail loudly. They will fail invisibly within systems that appear fully controlled. The institutions that build the semantic control plane now will not merely manage this risk. They will move faster than those that do not, because trusted infrastructure is foundational to speed at scale.

The introduction of agentic orchestration does not remove the need for control discipline. It raises the standard.

Institutions that fail to align control authority with execution architecture will not experience control failure as breakdown. They will experience it as valid execution of unintended outcomes.

References

ABA Banking Journal. (2025, December 8). Are We Sleepwalking Into an Agentic AI Crisis? American Bankers Association.

American Bankers Association and Bank Policy Institute. (2026, March). Joint Letter to NIST on Voluntary Guidance for Agentic AI Deployment in Financial Services.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331.

Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. (2011). Supervisory Guidance on Model Risk Management (SR 11-7). Federal Reserve.

Doyle-Spare, M. (2026, February 3). Super Bowl Strategy for Agentic AI: Context, Control, and Coordination. LinkedIn.

Doyle-Spare, M. (2026, February 10). Agentic AI the SMB Banking Advantage: SaaS as an Accelerator to Semantic Architecture. LinkedIn.

Doyle-Spare, M. (2026, March 10). Agentic Workflow Drift: A New Failure Mode in Banking Controls Testing. LinkedIn.

Doyle-Spare, M. (2026, forthcoming). Agentic Workflow Subversion: The First Enterprise Risk Born in the Reasoning Layer. LinkedIn.

Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR).

Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z., & He, C. (2024). LLM multi-agent systems: Challenges and open problems. arXiv preprint arXiv:2402.03578.

National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.

National Institute of Standards and Technology. (2026). AI Agent Standards Initiative: Request for Information on Securing Agentic AI Systems.

Shavit, Y., & Agarwal, S. (2023). Practices for governing agentic AI systems. OpenAI. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

Wang, C. L., Singhal, T., Kelkar, A., & Tuo, J. (2025). MI9 – Agent Intelligence Protocol: Runtime Governance for Agentic AI Systems. arXiv preprint arXiv:2508.03858.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. The Knowledge Engineering Review, 10(2), 115–152.

About the Author

Maureen Doyle-Spare is a senior executive operating at the convergence of technology, operations, and risk in global financial services. Her career spans global banking institutions, advisory and consulting, and technology services, where she has led P&L, built and scaled BFSI practices, and driven agentic AI transformation for some of the world's most complex regulated institutions. She brings practitioner experience across the full enterprise stack, from the control frameworks that govern banking operations to the agentic AI systems now being deployed inside them. This cross-domain perspective is what makes her vantage point distinct: she sees risk not from the model outward, but from the business and control boundary inward, grounded in direct experience leading transformation decisions and their downstream consequences. She is the originator of the Agentic Workflow Drift and Agentic Workflow Subversion frameworks and the Agentic 3 C's governance architecture — which is, to the author's knowledge, the first risk taxonomy developed specifically for the reasoning layer of agentic AI systems in financial services. This paper is submitted in an individual capacity as original practitioner-led research. Her ongoing work on agentic AI governance in financial services can be followed at linkedin.com/in/maureendoylespare.

Correspondence: linkedin.com/in/maureendoylespare

© 2026 Maureen Doyle-Spare. All rights reserved.
This working paper may not be reproduced or distributed without the author's permission.