SvedbergOpen
DISSEMINATIONOFKNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

Risk Dimensioning and Dispositioning: A Control-Based Production-Enablement Methodology for Governing Machine Learning, Generative, and Agentic AI in Financial Institutions

Sandeep Kumar Anuguthala

Independent Researcher, India. Email: anuguthalasandeepkumar@gmail.com

Abstract
Banks operate under regulatory obligations that require every risk introduced by a new system to be identified, measured, and formally accepted before that system touches customers, data, or decisions. Artificial intelligence is not exempt. It heightens the obligation, because AI systems behave probabilistically, drift over time, inherit bias from data, and can act autonomously in ways that are difficult to reverse. Without a formal gate, the production decision becomes a business decision made without a complete risk picture. This paper presents risk dimensioning and dispositioning, a control-based production-enablement methodology operated inside a first-line-of-defence technology risk and control function at a large financial institution between early 2024 and mid-2026. It takes a use case through six sequenced steps: intake and risk scoping, inherent risk identification, control posture validation, residual risk measurement and dispositioning, risk-informed decision facilitation, and non-objection and second-line approval facilitation. Dimensioning decomposes the system's risk profile across distinct domains and asks what exposure remains once controls are accounted for, rather than whether a control exists. Dispositioning closes the assessment into a documented production decision in which the use-case owner accepts residual risk, the second line of defence provides non-objection rather than co-approval, and Legal clears regulatory and contractual exposure, all three required before production, with the recurring discipline that the practitioner sights the artefact rather than accepting an assertion. The paper develops three of the framework's five risk dimensions, model risk, data risk, and technology controls, the third to the level of its control domains and verification standard. It reports the operating record of the programme the methodology ran inside, separately from the author's personal scope of 15 use cases led end-to-end through all six steps. No efficiency, speed, or cost claim is made, because no instrumented baseline exists.

Keywords: AI governance, machine learning risk management, generative AI governance, agentic AI risk, model risk management, first-line-of-defence, risk dimensioning, risk dispositioning, production-enablement, technology risk controls, financial institutions, AI regulatory compliance, residual risk assessment, control posture validation, enterprise AI governance, responsible AI deployment, operational risk, AI risk framework, pre-production risk gate, AI oversight

I. Introduction

The question a supervisor asks after an incident is always the same. Did you know the risk before you deployed, and if not, why not. A formal pre-production risk gate is an institution's answer to that question, given before production rather than constructed afterwards.

Financial institutions have spent two decades building the machinery to answer it for quantitative models, and for classical statistical and machine learning models it works reasonably well. Artificial intelligence has broken the fit. A generative assistant has no single model owner, no fixed training dataset under institutional control, and no traditional validation pathway. An agentic system chains decisions, calls external tools, and executes actions at machine speed. Neither is a model in the sense the existing frameworks assume, and neither stops being the institution's regulatory responsibility because of it.

The gap is not a shortage of frameworks. Institutions hold model risk management guidance, technology control frameworks, data risk frameworks, operational risk frameworks, and increasingly a general AI standard above them. The gap is that these frameworks are definitional: they state what categories of risk exist, not how to measure them for a specific use case, how to sequence or synthesise them, or how to reach a production decision from them. Compliance with a framework is not the same as having assessed the risk.

This paper describes a methodology built to close that gap and operated as a mandatory pre-production requirement. It is not a proposal. It was designed, iteratively refined, and executed inside a first-line-of-defence

(1st LOD) technology risk and control function at a large financial institution across approximately two and a half years of continuous practice, from roughly early 2024 through mid-2026, against use cases spanning machine learning, generative, and agentic configurations. The institution is not named, and every illustrative failure case in Section 12 is a composite rather than an account of a specific deployment.

The methodology has two halves the paper treats as one process. Dimensioning decomposes an AI system's risk profile across distinct domains, each measuring a different category of exposure the system introduces or heightens. Dispositioning converts a completed dimensional assessment into a formal, documented production decision. The two are inseparable: dimensioning without dispositioning produces a risk report nobody is obligated to act on, and dispositioning without dimensioning produces a decision nobody can defend. Three claims are made for the combination and no more, stated in Section 3.2, which also states what they do not extend to, including what is already claimed in the author's own prior published work.

2. Regulatory Context and Obligation

The obligation that makes a pre-production gate mandatory is not AI-specific. It is the general prudential requirement that a supervised institution identify, measure, and formally accept the risks a new system introduces before that system reaches customers, data, or decisions. In the United States it is expressed through the supervisory frameworks of the Office of the Comptroller of the Currency, the Federal Reserve, and the Federal Deposit Insurance Corporation, and through the institution's own board-approved risk frameworks operationalising them; for a large national bank the OCC's Guidelines Establishing Heightened Standards state it directly, requiring a written risk governance framework [12].

For anything functioning as a model, the anchor is the interagency supervisory guidance on model risk management, issued on 4 April 2011 as the attachment to Federal Reserve SR letter 11-7 and, jointly developed, as OCC Bulletin 2011-12 [6]. All quotations in this paragraph are from that dated 2011 document. It defines model risk as "the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports," and establishes four obligations the model risk dimension operationalises directly. The first is effective challenge, the guidance's own "guiding principle for managing model risk," defined as "critical analysis by objective, informed parties who can identify model limitations and assumptions and produce appropriate changes," carried out at a remove from the people who built the model: "Validation involves a degree of independence from model development and use," and "Independence is not an end in itself but rather helps ensure that incentives are aligned with the goals of model validation." The second is ongoing monitoring, which "is essential to evaluate whether changes in products, exposures, activities, clients, or market conditions necessitate adjustment, redevelopment, or replacement of the model." The third is a model inventory: "Banks should maintain a comprehensive set of information for models implemented for use, under development for implementation, or recently retired." The fourth is governance, because "model risk management encompasses governance and control mechanisms such as board and senior management oversight, policies and procedures, controls and compliance, and an appropriate incentive and organizational structure."
That guidance was revised during the period described here. On 17 April 2026 the Federal Reserve issued SR 26-2, which "supersedes and replaces SR letter 11-7, Guidance on Model Risk Management (issued April 4, 2011) and SR letter 21-8, Interagency Statement on Model Risk Management for Bank Systems Supporting Bank Secrecy Act/Anti-Money Laundering Compliance (issued April 9, 2021)," and the OCC issued Bulletin 2026-13, "Model Risk Management: Revised Guidance," which rescinds OCC Bulletin 2011-12 [7]. The practice reported in this paper was conducted against the 2011 guidance, the operative supervisory expectation throughout, and the four obligations quoted above are the ones the methodology was built on. No control requirement in this paper is drawn from the content of the 2026 revision.

One fact about the two documents is the sharpest available evidence for the argument this paper makes. The 2011 guidance does not mention artificial intelligence, machine learning, generative systems, or agentic systems anywhere across its twenty-one pages. The 2026 revision names generative and agentic AI once, in a footnote, and only to put them outside its reach: "Generative AI and agentic AI models are novel and rapidly evolving. As such, they are not within the scope of this guidance," with its principles stated to apply to "traditional statistical and quantitative models and non-generative, non-agentic AI models" [7]. Fifteen years and one full revision apart, the supervisory guidance that governs models in a bank either does not reach these systems or expressly declines to. The obligation to identify, measure, and formally accept their risk before production does not move out of scope with it, and closing that distance is what a methodology has to do.

Where AI outputs bear on consumer credit, a second body of obligation applies. The federal fair lending authorities, the Equal Credit Opportunity Act with its implementing Regulation B, and the Fair Housing Act, establish both the non-discrimination requirement and the adverse action requirement. Regulation B provides that a statement of reasons for adverse action "must be specific and indicate the principal reason(s) for the adverse action" (12 CFR 1002.9(b)(2)) [9]. That is why explainability is a control requirement rather than a reporting convenience in a regulated decision workflow: an institution that cannot state the specific principal reason for a decision cannot meet the obligation, whatever architecture produced it. The Bureau of Consumer Financial Protection's ECOA Baseline Review Modules set out how the risk is examined, directing examiners to "describe and evaluate the entity's fair lending compliance management system (CMS), including board and management oversight and compliance program (policies and procedures, training, monitoring and/or audit and consumer complaint response)," and to describe how the entity manages the fair lending risks posed by its service providers [9]. Those are the document's own terms, and the assessment record is built to answer them in them.

Data obligations arrive from three directions. The Gramm-Leach-Bliley Act, through the Interagency Guidelines Establishing Information Security Standards, requires each institution to "implement a comprehensive written information security program that includes administrative, technical, and physical safeguards" for customer information. The specific measures are framed differently, and the difference matters for how a control is evidenced: the institution "must consider whether the following security measures are appropriate" and "adopt those measures the [institution] concludes are appropriate," the first being "access controls on customer information systems, including controls to authenticate and permit access only to authorized individuals" [10]. The programme is a flat requirement; the access control is a measure to be considered and adopted where appropriate, so the assessment record has to evidence the reasoning as well as the control. Oversight of service provider arrangements is a separate standard in the same Guidelines, elaborated for a third-party AI service by the interagency guidance on third-party relationships [11]. The General Data Protection Regulation imposes purpose limitation, personal data being "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes," storage limitation, and a right to erasure under Article 17; and the California Consumer Privacy Act gives a consumer "the right to request that a business delete any personal information about the consumer which the business has collected from the consumer" [13]. Each is a live control obligation for an AI system that ingests, retrieves, or writes customer data. The NIST AI Risk Management Framework and its Generative AI Profile [8] sit alongside all of these as voluntary standards organising AI risk into functions, categories, and generative-specific risk classes. They are a useful common vocabulary, and they are definitional in the way Section 4 describes. They do not produce a production decision.

Two things follow. First, AI escapes none of these obligations. It heightens them, because probabilistic behaviour, drift, inherited bias, and autonomous action each make the obligation harder to discharge rather than less applicable. Second, and counterintuitively, the argument of this paper is that a functioning production gate accelerates AI adoption rather than slowing it: when the process is trusted, risk issues surface upstream during design rather than at the gate, late-stage blockers reduce, and the institution can demonstrate with evidence that its deployment posture is mature. That is the argued mechanism, stated as an argument. It is not a measured result. No instrumented baseline exists from before the methodology was in place, so this paper makes no comparative claim about speed, cost, or throughput.

3. Related Work and Contribution Boundary

Published work on AI governance clusters into three groups, none of which occupies the layer this paper addresses.

The first is standards and supervisory guidance. The 2011 interagency model risk guidance [6] specifies what sound model risk management looks like, the NIST AI Risk Management Framework and its Generative AI Profile [8] specify which categories of AI risk an organisation should govern, and the 2023 interagency guidance on third-party relationships [11] specifies how a banking organisation should manage risk across the life cycle of a vendor relationship. All are framework documents by design rather than assessment methodologies.

The second is technical research on the failure modes that make AI hard to govern. Gama et al. [14] survey concept drift detection, establishing the measurement problem for models whose input distribution moves after deployment. Fuster et al. [15] show, in the credit setting specifically, that machine learning can change the distribution of outcomes across borrower groups, the empirical basis for treating bias testing as a load-bearing control rather than a documentation exercise. Perez et al. [16] show that automated red-teaming surfaces failure modes standard evaluation does not reach, and Greshake et al. [17] demonstrate indirect prompt

injection against real language-model-integrated applications, precisely the failure the technology controls dimension is scoped to catch. Chan et al. [18] characterise the harms arising as algorithmic systems become more agentic, including consequences propagating through actions rather than outputs. This literature defines the risks well. It does not address how a regulated institution converts knowledge of them into a production decision.

The third is governance process research. Raji et al. [19] set out an end-to-end internal algorithmic auditing framework, the closest published work to the evidence discipline of Section 10 and sharing its premise that assurance requires artefacts rather than assertions. Davies and Zhivitskaya [20] examine the three-lines-of-defence model critically, including the question Section 11.2 answers: who owns a risk decision when a second line is involved.What the signal must do.

3.1 Relationship to the author's prior published work

The author's prior work occupies adjacent layers of the same programme, and the boundary needs stating precisely rather than left to inference.

Enterprise Agentic AI Lifecycle Governance [1] establishes a control-driven governance framework spanning the agentic AI lifecycle from planning and design through deployment, monitoring, and decommissioning, including inherent risk assessment, risk tiering, and continuous control validation. That paper is organised along the lifecycle axis: which governance activity applies at which stage. The present paper is organised along the dimensional axis: which categories of risk apply to a given use case, and how they resolve into one decision at a single point in that lifecycle, the production gate. The lifecycle contribution is not restated here.

That axis distinction is real at the level of framing and not sufficient on its own, so two specific overlaps with [1] are stated rather than left to inference. First, at step level: Step 2 of Section 5, inherent risk identification, and Step 3, control posture validation, cover the same substance as [1]'s inherent risk assessment and continuous control validation. What differs is the unit and the moment. In [1] those activities are stage-gated governance distributed across a system's life and applied to the system as a whole; here they are applied per risk dimension to a single use case, with the applicability scoping of Step 1 ahead of them and a per-dimension residual synthesis after them, resolving into one decision at one gate. Neither the activities nor their lifecycle placement is claimed as novel here. Second, at rating level: the low, moderate, high, critical residual rating of Section 11.1 is adjacent to [1]'s risk tiering, which that paper did not distinguish from it. Tiering classifies a system's overall risk level and functions as an input determining which governance activities apply; the rating here is an output of the completed dimensional assessment, carrying defined dispositioning implications at the gate rather than selecting activities upstream.

GUARD [2] specifies runtime monitoring and incident response for enterprise agentic AI, extending the lifecycle framework into post-deployment enforcement. It begins where this paper ends: the present methodology terminates at production clearance and claims nothing about runtime enforcement, detection, or incident response. Preventing Shadow AI Workloads [4] and Shadow MCP [5] address the predecessor problem, knowing which AI workloads and tool servers exist at all, through registry, build-time enforcement, and discovery. That problem is logically prior: a pre-production gate governs what is brought to it and cannot govern what is never declared, which matters for the reason recorded in Section 13.2, that the single internal audit finding over the period concerned inventory completeness rather than the assessment methodology.

Enterprise AI Supply Chain Security Governance [3] introduces an AI Bill of Materials and a weighted risk score for adopting open-source AI models, and disposes of individual model artefacts into risk-rated outcomes including conditional approval and rejection. This is the closest adjacency to the present paper's third claim, and it is named rather than glossed. Three distinctions hold. The unit of assessment differs: [3] assesses a model artefact entering the estate, this paper assesses a use case seeking production. The synthesis mechanism differs: [3] aggregates category scores arithmetically, whereas the methodology here rejects arithmetic aggregation and treats synthesis as a documented practitioner judgement (Sections 6.2, 11.1). That second one is a difference in design rather than a contribution, and Section 3.2 does not claim it as one. The distinction that carries the third claim is the third: the allocation of decision rights set out in Section 11.2 is not specified in [3], nor in [1] or [2]. The approve, approve-with-conditions, reject close already exists in [3], so here it is the vehicle that carries the allocation rather than the claim.

3.2 What is and is not claimed as novel

Three things are claimed, and nothing beyond them.

1. The integration layer, claimed as the integration itself and nothing more. The risk domains are assessed as an interacting set that resolves into one documented residual risk picture at a single gate, with the cross-dimensional interactions recorded explicitly in that record rather than left implicit across independently completed assessments. Section 6.2 works the interactions the record has to carry. The claim is the integration and its documentation. It is not a claim to a novel synthesis rule: the methodology deliberately specifies no arithmetic aggregation rule, and that absence is a design choice, not a contribution.

2. The formalised scoping judgement. The determination of which risk scenarios from the institution's frameworks are materially applicable to a specific use case, and whether the use case introduces net-new risk or heightens existing exposure, is named as a discrete, documented, auditable step.

3. The allocation of decision rights at the production gate. The use-case owner accepts the residual risk, the first-line practitioner facilitates and documents but never decides, the second line of defence gives non-objection rather than co-approval, and Legal clears regulatory and contractual exposure, with all three required before production. That allocation is the claim. The sequence that carries residual risk measurement through control gap structuring to an approve, approve-with-conditions, or reject close is the vehicle for it, and is not claimed as novel, because a three-outcome close of that kind already exists in the author's prior published work [3].

The individual risk domains are not claimed as novel. Model risk, data risk, technology controls, and operational risk are established domains with mature literatures and, in a large institution, mature internal frameworks. The methodology does not replace those frameworks; it sits above them and orchestrates them into a decision. No claim is made to novelty in drift detection, bias measurement, red-teaming technique, prompt injection defence, or agentic safety research. The methodology consumes that work as control requirements rather than advancing it.

4. Why Current Approaches Fail

Most financial institutions approach AI production readiness through one of three patterns. None is sufficient alone, and the reasons are structural rather than a matter of execution quality.

Table 1. Current-state patterns for AI production readiness and the failure mode of each.

Pattern	What the institution does	Why it is insufficient
Framework compliance	Adopts SR 11-7 for model risk and a general AI standard such as the NIST AI RMF or ISO/IEC 42001, then reviews each use case against it as a pre-production checklist	The frameworks are definitional: what categories of risk exist, not how to measure, sequence, or decide from them (Section 1)
Model risk management alone	Routes every AI system through the existing MRM process	MRM was designed for classical statistical and machine learning models. It does not fit generative systems with no single model owner, no fixed institutionally controlled training set, and outputs that cannot be validated by backtesting. It fits agentic systems less still, because those act autonomously, chain decisions, and call external tools and APIs, introducing operational and technology risk dimensions MRM was never built to capture
Ad-hoc review	A committee, working group, or senior risk officer reviews the use case, asks questions from experience, and makes a judgement call	Better than nothing, but not repeatable, not consistent across reviewers, not auditable in any meaningful sense, and not scalable as volume grows. Two reviewers assessing the same use case can reach materially different conclusions and neither can fully explain why

The deeper problem is that the three AI paradigms carry structurally different risk profiles, and no single-methodology approach handles all three coherently. Machine learning systems carry model risk as their primary dimension, bias, drift, validation, and performance degradation, which existing frameworks handle reasonably well while underweighting technology controls and data provenance as co-equal dimensions. Generative systems introduce risks model risk management was not built for: hallucination, prompt injection, output unpredictability, third-party model dependency, data leakage through retrieval layers, and the absence of a traditional validation pathway. Checklist approaches tend to bolt generative AI onto an existing model risk checklist with a few extra questions, creating a false sense of coverage without assessment depth. Agentic systems are the most exposed, combining the model risk of machine learning with the output unpredictability of generative AI and adding an entirely new class of exposure: autonomous action, tool use, decision chaining, and errors propagating at machine speed before any human can intervene. Governed through one ad-hoc or checklist-based process, the review defaults in practice to whichever dimension the reviewer knows best, usually model risk, and the rest receive superficial treatment: a production decision compliant in appearance and incomplete in substance.

5. The Methodology: Six Steps from Intake to Non-Objection

The methodology takes a use case through six sequenced steps before production clearance. Figure 1 shows the sequence and the three-party gate that terminates it.

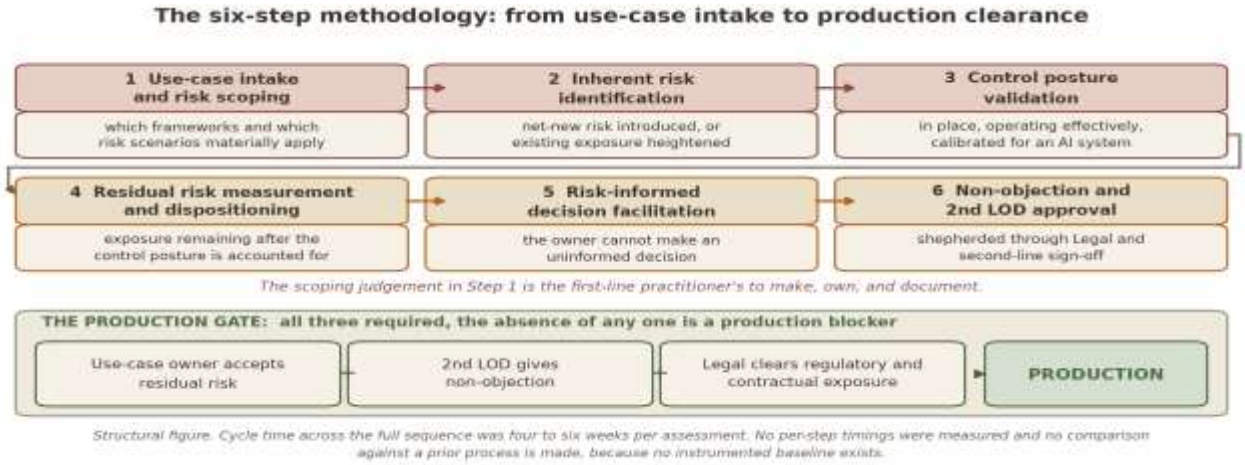


Figure 1. The six-step sequence from use-case intake to production clearance, with the three-party approval gate. The first-line practitioner facilitates and documents; the decision itself is allocated as Section 11.2 sets out.

Step 1: use-case intake and risk scoping. The practitioner sits with the use-case owner and establishes what the system does, how it is built, what data it touches, and who it affects. From that comes the first critical judgement: which risk frameworks, and which specific risk scenarios within them, are applicable. The methodology operationalises requirements drawn from four institutional risk frameworks, Model Risk Management (MRM), the Technology Control Framework (TCF), Data Risk Management (DRM), and the broader Operational Risk Management (ORM) framework, consistent with the institution's AI governance standard for the safe and responsible development, deployment, and use of AI and machine learning systems. Within the TCF, applicable control domains include Information and Technology Asset Management, Identity and Access Management (IAM), Cryptography, Vulnerability Management, Technology Development, and Technology Resiliency.

The scoping decision belongs to the first-line practitioner, because not every scenario applies universally. The practitioner assesses whether the use case introduces net-new risk to the firm or heightens an existing exposure, and which controls are genuinely load-bearing for this deployment rather than nominally applicable. That separation of applicable from inapplicable risk scenarios is where the methodology's analytical rigour sits, and it is the step ad-hoc processes leave tacit. Formalising it as a named, documented step is what makes it auditable, and what the methodology relies on to make it transferable rather than resident in one practitioner's judgement. Auditability follows from the record; transferability is a design intent this paper does not evidence, and Section 14 records why.

Step 2: inherent risk identification. Across each applicable dimension, the practitioner identifies the raw, unmitigated risks the use case introduces or heightens. This is not a checklist exercise: it requires establishing whether the use case brings net-new risk or amplifies existing exposure, because the two carry different control implications and different escalation paths.

Step 3: control posture validation. The practitioner assesses whether controls are in place, designed and operating effectively, and calibrated for an AI or machine learning system rather than a generic technology system, spanning IAM, vulnerability management, data governance, model validation, cryptography, technology resiliency, and technology development. The verification discipline that makes this step meaningful rather than documentary is the subject of Section 10.

Step 4: residual risk measurement and dispositioning. After accounting for the control posture, the practitioner measures residual risk and works with the owner to establish what the rating means in practice: what it allows, what it constrains, and what the owner is accepting if they proceed (Section 11).

Step 5: risk-informed decision facilitation. The practitioner does not make the go or no-go decision. That belongs to the use-case owner. The obligation is to structure the information so the owner cannot make an uninformed decision.

Step 6: non-objection and second-line approval facilitation. The practitioner shepherds the use case through Legal review and second-line sign-off. Nothing moves to production without those cleared.

The methodology therefore sits inside a three-lines-of-defence structure and takes a position on a question that model has long left ambiguous [20]: what the second line owns when it signs.

6. Risk Dimensioning

Risk dimensioning is the structured decomposition of an AI system's risk profile across distinct, non-overlapping domains, each measuring a different category of exposure the system introduces or heightens. The purpose is not to produce an aggregate score prematurely, but to ensure every material risk category has been examined on its own terms before any cross-dimensional synthesis. It differs from a checklist in one critical way. A checklist asks whether a control exists. Dimensioning asks what the underlying exposure is, whether the control is adequate for that specific exposure, and what residual risk remains once the control is accounted for. That distinction is the difference between compliance theatre and genuine risk assessment.

6.1 Scope of this paper: three dimensions of five

The framework comprises five risk dimensions, each producing its own residual risk assessment before cross-dimensional synthesis (Section 11.1).

This paper develops three of the five in depth: Dimension 1, Model Risk (Section 7); Dimension 2, Data Risk (Section 8); and Dimension 3, Technology Controls (Section 9). The remaining two are part of the methodology as operated, and are identified as such, but they are reserved for separate treatment rather than specified here.

This is deliberate scoping, not an omission. Each of the three developed here carries a full control specification, a verification standard, and a distinct regulatory anchor. Three at that depth are more useful to a practitioner than five at summary depth, and presenting five thinly would reproduce the failure mode Section 4 criticises, risk categories enumerated without the measurement and verification detail that makes them operable. Section 14 records the consequence. Figure 2 shows how the three behave differently across machine learning, generative, and agentic systems, recording structural differences in what must be assessed rather than measured incidence or severity.

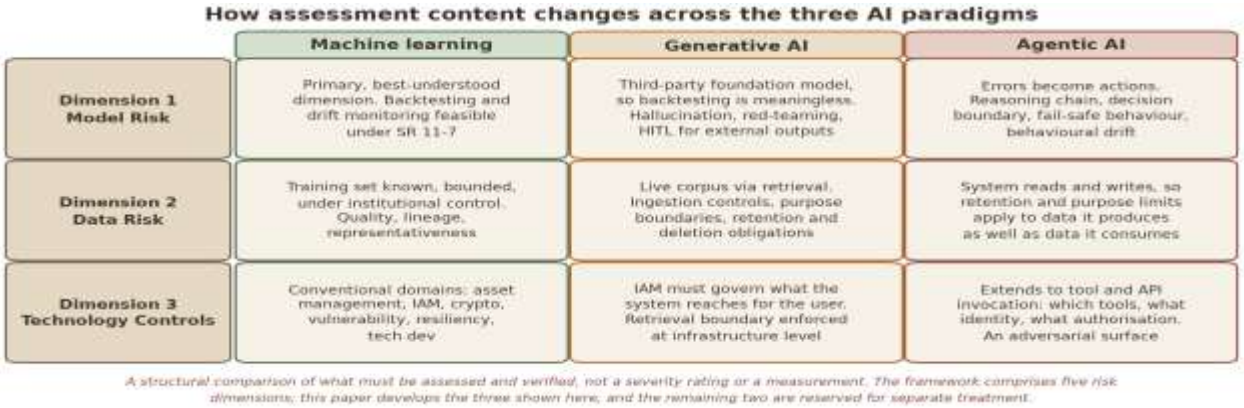


Figure 2. How the assessment content of each dimension changes across the three AI paradigms. This is a structural comparison of what must be assessed and verified, not a measurement or a severity rating.

6.2 The integration layer, worked

Assessing dimensions separately is the easy half. The claim in Section 3.2 is about what happens between them, so three interactions are worked here rather than asserted, each drawn from Sections 7 to 9 and each stated as the analytical structure of the method rather than a measured effect.

Data risk conditions the evidentiary value of model risk. For a model bearing on a credit decision, bias and fairness testing is a Dimension 1 control and lineage documentation is a Dimension 2 control, and the second determines what the first is worth. A bias test run against a dataset whose provenance and completeness are undocumented is evidence that a computation was performed, not that the model is fair on the population it will be applied to. A high residual in data risk therefore does not simply sit beside a moderate residual in model risk; it degrades the standing of the model risk evidence. This is the concrete reason the methodology refuses arithmetic aggregation: averaging a high against a moderate returns a number that conceals a one-directional dependency.

Technology controls are the delivery mechanism for a data risk obligation. For a retrieval-based generative system, the purpose boundary approved under Dimension 2, evidenced by a dated and attributed Data Use Council approval, is enforced in production by a Dimension 3 control, the retrieval-layer access boundary scoped to the querying user's authorisation. If that boundary is not enforced, the approval record describes a purpose the running system does not observe, so the Dimension 2 residual is conditional on a sighted Dimension 3 configuration rather than closable on the approval alone. The generative composite in Section 12 is that link failing, which is why it presents as a technology control failure and a data governance failure at once.

One regulatory obligation can be carried by three dimensions at once. Regulation B's requirement that a statement of reasons be specific and indicate the principal reasons [9] is discharged operationally by an explainability control in Dimension 1. But an explanation has to be reconstructable after the fact, which depends on Dimension 3 logging and configuration, and attributable to inputs, which depends on Dimension 2 lineage. None can be signed off in isolation without leaving the obligation partly unevidenced, and a per-dimension review that never looks across the set records three acceptable ratings and one undischarged obligation.

The pattern is directional: where one dimension supplies the mechanism by which another's control is delivered, the dependent dimension's residual cannot honestly be rated better than the mechanism's. That is an ordering the practitioner judgement has to respect, deliberately not expressed as a formula. What the methodology requires instead is that the residual record state the dependency and not only the two ratings, so a later reviewer sees the link rather than reconstructing it. That documented dependency is the whole of the integration claim. Two boundaries on it: no aggregation rule is specified or claimed, and a five-dimension framework contains ten dimension pairs, of which three are worked here.

7. Dimension 1: Model Risk

What it measures. The risk that the system produces incorrect, biased, or unreliable outputs, and that those outputs drive decisions or actions with material consequences for the firm or its customers. The dimension examines model development rigour, validation independence, performance metrics, bias and fairness testing,

ongoing monitoring, and drift over time. It asks whether the model does what it claims to do, whether that claim has been independently verified, and whether the institution will know when performance degrades.

7.1 Human oversight as a first-order control

Bias risk is materially heightened when AI outputs flow into critical workflows that touch external parties. In those contexts, human-in-the-loop (HITL) design is a first-order control requirement rather than an optional enhancement. The reasoning is about reversibility: a biased or inaccurate output reaching an external stakeholder carries regulatory, reputational, and customer-harm implications that cannot be remediated after the fact. A control that can only be applied before the output leaves the institution is load-bearing, not a maturity improvement to be scheduled later.

7.2 Adversarial testing as a mandatory control

For generative and agentic systems, adversarial testing and AI red-teaming are mandatory model risk controls, because traditional backtesting is structurally insufficient. Red-teaming surfaces failure modes standard evaluation frameworks do not reach: hallucination under adversarial prompting, reasoning chain manipulation, and out-of-bound outputs. The literature supports this directly. Perez et al. [16] show that systematic adversarial probing uncovers harmful behaviours ordinary evaluation misses, and Greshake et al. [17] show that adversarial content delivered through an application's own retrieval path can subvert its behaviour. The NIST Generative AI Profile treats this class of exposure as a set of generative-specific risk categories requiring their own measurement rather than as a variation on model performance [8]. A control set that assumes benign inputs is not calibrated for these systems.

7.3 Behavioural drift is not performance drift

For agentic systems, model risk extends beyond output quality into action risk. The agent's reasoning chain must be subject to explainability controls, behavioural validation, and recalibration mechanisms that detect and address behavioural drift promptly.

This is a distinct exposure from traditional model drift, and the distinction is operationally important. Concept and performance drift are well characterised and detectable through monitoring of input distributions and output accuracy against defined thresholds [14]. Behavioural drift is directional change in what the system does rather than degradation in how accurately it does it, and may not be visible in standard output metrics at all: an agent that begins resolving a class of ambiguous cases differently has changed behaviour without changing any monitored performance statistic. The control requirement therefore cannot be satisfied by a performance threshold alone.

7.4 How the dimension differs across paradigms

Machine learning. Model risk is the primary and best understood dimension. SR 11-7 provides a mature framework [6], validation through backtesting, benchmarking, and out-of-sample testing is feasible, and drift monitoring is operationally achievable [14]. The risk is real but manageable through established practice.

Generative AI. Traditional model risk management breaks down structurally. The institution typically relies on a foundation model developed and maintained by a third party, which makes independent validation through backtesting meaningless. That dependency moves part of the control burden off model validation and onto third-party oversight, where the operative supervisory expectation is the interagency guidance on third-party relationships, covering risk management across the full life cycle of the relationship from due diligence through contracting, monitoring, and termination [11]. Bias manifests differently, not as statistical disparity in a prediction but as systematic tendencies in generated content that are harder to measure and to remediate. Hallucination introduces a category of model risk, confident incorrect output, with no direct analogue in classical machine learning. Where outputs feed critical or externally facing workflows, bias becomes a primary risk driver and HITL a mandatory control response.

Explainability requires particular care. When a generative system contributes to a credit, fraud, or customer-facing decision in a regulated context, Regulation B's requirement that a statement of reasons be specific and indicate the principal reasons [9] admits no architectural exemption, so black-box output is not a defensible posture. Explainability is a control validated before production clearance, confirming that outputs can be interrogated, attributed, and explained at a level of specificity the institution can stand behind. Model risk for generative systems is accordingly assessed through output testing, red-teaming, human evaluation protocols, and third-party model dependency risk rather than through backtesting.

Agentic AI. Model risk compounds, because errors do not stay inside the model. They propagate into actions. A wrong prediction from a machine learning model is a bad output; a wrong decision from an agentic system triggers a downstream action that may be irreversible, and the harms characteristic of increasingly agentic systems follow from that property [18]. Model risk here must include reasoning chain reliability, decision boundary validation, behavioural drift detection and recalibration, and failure mode behaviour. The operative question is what the agent does outside its training distribution: does it fail safely, or does it interpolate. Explainability is a control mechanism rather than a reporting convenience, because without it the reasoning chain that produced an action cannot be reconstructed after the fact.

8. Dimension 2: Data Risk

What it measures. The risk arising from the data the system consumes: its provenance, quality, completeness, representativeness, and lineage, and the controls governing its access, handling, and retention. The dimension examines whether the data used to train, fine-tune, or operate the system is fit for purpose, whether it carries embedded bias or gaps that will propagate into outputs, whether data governance controls are operating effectively, and whether the system's data handling creates regulatory exposure under privacy, fair lending, and jurisdictional data use obligations. Data risk is treated as co-equal with model risk rather than supporting detail, one of the specific underweightings identified in Section 4.

Data quality and ingestion controls are a primary focus, because inadequately curated data increases the potential for unfit-for-purpose outputs regardless of how well the model is built. The ingestion pipeline is validated on three points: what content is ingested, how it is quality-checked before use, and whether ingestion controls prevent low-quality, outdated, or inappropriate material from entering the knowledge base. For a retrieval-based generative system that is not a training-time concern closed once. It is a continuing control on a live corpus.

8.1 Data use governance and purpose boundaries

Data use governance is a distinct control requirement, separate from data quality. Data sourced for a specific use case and business purpose must not be used for additional purposes contrary to jurisdictional or regulatory obligations. Data Use Council approval and data governance sign-off are mandatory controls, not only for quality but for purpose boundary enforcement, because unauthorised secondary use creates both regulatory exposure and contractual dispute risk with data providers. The anchor is direct: GDPR requires that personal data be collected for specified, explicit, and legitimate purposes and not further processed incompatibly with those purposes [13].

8.2 Retention and deletion as a distinct exposure

Retention and deletion obligations require explicit assessment, independent of both data quality and data use governance. When a generative system processes customer data through a retrieval layer, or an agentic system reads and writes data as part of its workflow, how long that data is retained, where it is stored, and whether deletion obligations can be met are live compliance exposures. GDPR Article 17 establishes a right to erasure, GDPR storage limitation requires that personal data be kept in identifiable form no longer than necessary, and the CCPA establishes a consumer right to request deletion [13]. A system can satisfy every input-side quality control while failing these entirely, and they must be confirmed before production rather than deferred to a privacy review running on a different clock.

8.3 How the dimension differs across paradigms

Machine learning. Data risk centres on training data quality, lineage, and representativeness. The dataset is typically known, bounded, and under institutional control, and the key questions are whether it is complete, clean, and representative of the population the model will be applied to. Data bias at the input level produces biased outputs, and in a credit context that is the mechanism by which a technically well-performing model creates fair lending exposure [15]. That is why the data layer, not only the model layer, is where bias testing obligations attach.

Generative and agentic systems. The controlled-dataset assumption does not hold. The exposure moves from a bounded training set to a live corpus reached through a retrieval layer, and from read-only consumption to read-and-write operation. That relocation makes ingestion controls, purpose boundaries, and retention and deletion obligations the operative data risk controls for these paradigms rather than training-set representativeness alone.

9. Dimension 3: Technology Controls

What it measures. Whether the system satisfies the institution's existing technology control requirements, calibrated for an AI system rather than a generic technology deployment. This is the broadest dimension and the one mapping most directly to the existing technology risk framework. Its content is the applicable control domains of the Technology Control Framework, including but not limited to Information and Technology Asset Management, Identity and Access Management, Cryptography, Vulnerability Management, Technology Development, and Technology Resiliency. It is not a new control library. It is the determination of which existing controls are load-bearing for this deployment and whether they hold.

Two features distinguish it from a conventional technology risk review. The first is that AI systems change what several of these domains have to cover. Identity and access management for a conventional application governs who can use the application; for a retrieval-based generative system it must additionally govern what the system can reach on the user's behalf, so the access boundary has to be enforced at the retrieval layer and scoped to the querying user's authorisation, not merely at the interface. This is the control gap in the generative composite of Section 12, and the mechanism by which content becomes reachable through an AI interface that the same user could not access directly. For an agentic system the question extends to tool and API invocation: what the agent can call, with what identity, under what authorisation. The underlying obligation is not novel. The Interagency Guidelines Establishing Information Security Standards place "access controls on customer information systems, including controls to authenticate and permit access only to authorized individuals" among the measures an institution must consider and, where appropriate, adopt, inside a written information security programme the Guidelines do require outright; and they make service provider oversight a standard in its own right [10], elaborated for a third-party AI service by the third-party relationships guidance [11]. What is new is where the boundary now has to be enforced and demonstrated.

The second is that the retrieval and tool-use surface is an adversarial surface. Content reached through a retrieval layer can carry instructions that subvert the system's behaviour [17], which makes content filtering and retrieval boundary controls security controls rather than quality controls. The verification standard of Section 10 applies here with particular force: a design document stating that access is restricted to authorised data stores is not evidence that access is restricted. A confirmed technical configuration is.

This paper presents Dimension 3 at the level of its control domains, their AI-specific recalibration, and the verification standard. The per-domain control and evidence specification is not exhaustively enumerated, and Section 14 records that boundary.

10. Controls, Evidence Standards, and Sighting the Artefact

Most AI governance frameworks fail at the same point. They identify risk categories, list control objectives, and stop. A questionnaire is completed, a committee reviews it, the system moves to production, and nobody has verified that a single control is in place, operating, and calibrated for an AI system. The resulting record is internally consistent and operationally meaningless: it documents what should be true rather than what is true, so when something goes wrong it reflects what the owner believed or was told, not the control environment that existed at production. This is the gap the internal algorithmic auditing literature identifies from the assurance side [19]: an audit that consumes claims rather than artefacts cannot produce assurance. The methodology closes it through one recurring discipline**. The practitioner sights the artefact.** For every control asserted there is a defined evidence standard, and the assertion is not accepted in place of the evidence.

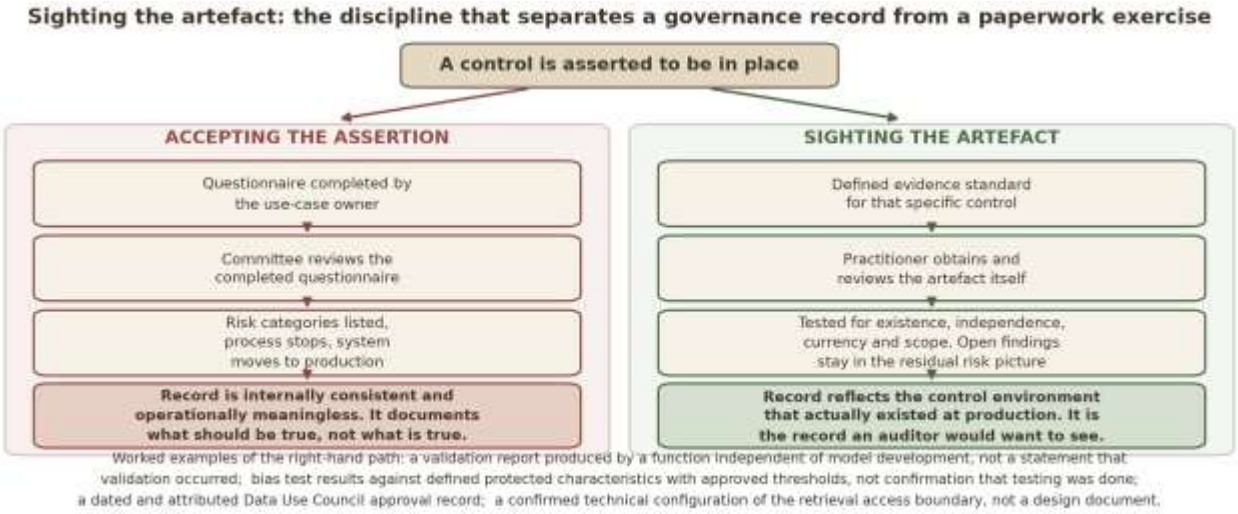


Figure 3. The control verification pattern. The left path is the failure mode the methodology is built to prevent; the right path is the discipline that replaces it. The distinction is not the strength of the control but whether its operation was evidenced.

Table 2 sets out how the three dimensions map to concrete controls, the evidence standard by which each is verified, and the existing regulatory obligation it discharges. The last column matters: the methodology creates no new regulatory requirements. It provides the operational mechanism for demonstrating compliance with requirements the institution already has.

Table 2. Control, evidence, and regulatory mapping for the three dimensions developed in this paper.

Dimension	Concrete controls the system must satisfy	Evidence standard: what the practitioner sights	Regulatory anchor
1. Model Risk	Independent validation; documented development methodology; performance benchmarking against defined thresholds; bias and fairness testing against applicable protected characteristics; ongoing monitoring with defined triggers; model inventory registration. Generative: output quality testing, red-teaming, HITL design documentation, explainability validation for regulated decision contexts. Agentic: reasoning chain validation, behavioural boundary testing, drift detection design	A validation report that exists, was produced independently of model development, addresses the dimensions relevant to this use case, and has been accepted by the model risk function. Open findings remain in the residual picture; they do not disappear because validation was technically completed. Bias testing: documented results against defined protected characteristics with thresholds approved by the model risk function, not a statement that testing was performed. Generative: red-team documentation naming who ran the exercise, which scenarios were tested, which failure modes surfaced, and how each was addressed	The April 4, 2011 interagency model risk guidance [6]: effective challenge with a degree of independence from development and use, ongoing monitoring, model inventory, governance and control mechanisms. ECOA and Regulation B, Fair Housing Act, and the CFPB ECOA Baseline Review Modules [9] for bias testing and explainability in consumer-facing decisions. Third-party relationships guidance [11] for a third-party foundation model

2. Data Risk	Input layer: data quality validation, lineage documentation, classification, ingestion controls confirming that only curated, classified, authorised data enters the knowledge base. Governance layer: Data Use Council approval for the approved purpose, data governance sign-off on cross-jurisdictional use, privacy impact assessment where personal data is processed. Retention layer: defined retention periods, deletion mechanisms, confirmation that jurisdictional deletion obligations can be fulfilled. Retrieval-based generative: retrieval boundary validation and content filtering	Lineage documentation as a concrete artefact tracing data from source through transformation into the system, reviewed for completeness, currency, and consistency with what the owner described; incomplete lineage is a control gap, not an acceptable state. Data Use Council approval verified through the record itself, dated and attributed, not the owner's representation that it was obtained. Privacy impact assessment verified through the completed document and the resolution of its findings before production. Retrieval boundary: a technical demonstration or architecture review confirming enforcement at the infrastructure level	GLBA Interagency Guidelines: a required written programme, with the specific measures on a consider-and-adopt basis [10]. GDPR purpose limitation, storage limitation, and right to erasure, and the CCPA deletion right [13]. Fair lending authorities at the data layer [9]. The institution's data governance framework, which establishes the Data Use Council requirement and classification obligations
3. Technology Controls	Applicable domains of the Technology Control Framework, including Information and Technology Asset Management, Identity and Access Management, Cryptography, Vulnerability Management, Technology Development, and Technology Resiliency, calibrated for an AI system. Retrieval-based and agentic: access boundary enforcement at the retrieval layer scoped to the querying user's authorisation, and control over which tools and APIs an agent may invoke, with what identity and under what authorisation	Confirmed technical configuration rather than design intent. A design document stating that access is restricted to authorised data stores is not evidence that access is restricted; a confirmed configuration is. Per-domain evidence specifications are not exhaustively enumerated (Section 14)	GLBA Interagency Guidelines: access controls on customer information systems as a measure to be considered and adopted where appropriate, and service provider oversight as a standard in its own right [10]. Third-party relationships guidance for tool, API, and managed-service dependencies [11]. The institution's Technology Control Framework

Because every disposition rests on sighted artefacts rather than attestations, the record the assessment produces is the same record an internal auditor or a supervisor would ask for, not a parallel document constructed for review. That is what makes the output auditable in the sense the ad-hoc pattern of Table 1 cannot be.

11. Dispositioning and Decision Ownership

Dispositioning converts a completed dimensional risk assessment into a formal production decision. If dimensioning answers what risk this system carries, dispositioning answers what the institution does about it. It is not a rubber stamp, and it is not the risk function telling the business whether it can proceed. It is a

structured facilitation in which the practitioner presents a complete, evidence-based residual risk picture and the owner makes a documented, risk-informed decision within their delegated authority.

11.1 The dispositioning sequence

Step 1: residual risk determination. Residual risk is what remains after the current control posture has been accounted for. It is not the inherent risk of the use case, and it is not the control inventory. It is the net exposure the institution carries if the system goes to production in its current state.

It is determined dimension by dimension first, then synthesised into an overall rating. The synthesis is not a mathematical average. It is a practitioner judgement accounting for the interactions Section 6.2 works, and it is the point on which the methodology differs from weighted-score approaches, including the author's own supply chain scoring work [3]. The rating sits on a defined scale, typically low, moderate, high, or critical, each level carrying defined dispositioning implications. The rating is not the decision. It is the evidence base from which the decision is made.

Step 2: control gap identification and mitigation structuring. Where residual risk is elevated, because controls are absent, not operating effectively, or inadequately calibrated for an AI system, the practitioner identifies the specific gaps driving the elevation and works with the owner to determine whether they can be closed before production, closed after production under a defined remediation commitment, or accepted within the owner's delegated authority. Closing all material gaps before production opens the path to unqualified approval; gaps persisting post-production under defined mitigations and timelines lead to approval with conditions; and where no reasonable mitigation brings residual risk within acceptable bounds, rejection or deferral is the appropriate outcome.

The practitioner does not make this determination unilaterally. Their role is to present the gap picture with precision: what is missing, why it matters, what the residual exposure is without it, and what a credible mitigation looks like. The owner engages with that picture and determines what they will commit to.

Step 3: the disposition decision. Three outcomes are available, and they are the vehicle for the decision-rights allocation of Section 11.2 rather than a contribution in themselves (Section 3.2).

Approved. Inherent risks have been identified across all applicable dimensions, controls are in place and operating effectively, and residual risk falls within the institution's appetite for deployments of this type. The owner formally accepts the residual risk and the system is cleared, and the acceptance is documented: not an email or a verbal sign-off, but a recorded, attributed decision identifying the decision-maker, the rating accepted, and the date.

Approved with conditions. Control gaps are real but addressable. The owner proceeds under documented conditions: controls to be implemented within defined timeframes post-production, monitoring from day one, HITL requirements for specific output types, or operational constraints limiting scope until mitigations are confirmed. This is not a weaker approval. It creates a formal accountability structure for closing the gaps, and the conditions are tracked, owned, and validated rather than filed in a governance document nobody revisits.

Rejected or deferred. Residual risk cannot be adequately mitigated within the proposed configuration. Rejection means the use case cannot proceed as currently designed, not that it cannot proceed: the owner receives an articulation of what would have to change in design, data handling, control environment, or operational scope for the profile to become approvable. Deferral applies where the owner needs time to implement those changes before resubmitting.

11.2 Who makes the call

This is one of the most consequential governance design questions in the methodology, and one published frameworks answer incorrectly or not at all [20].

The disposition decision belongs to the use-case owner. Not the risk function, not the second line, not the technology team. The owner holds delegated business authority over the system, benefits from it reaching production, and carries the accountability if something goes wrong after it does. The decision must sit with that accountability.

The first-line risk practitioner does not approve or reject. They dimension, assess, structure the control gap picture, facilitate the dispositioning conversation, and document the outcome. The role is to ensure the owner cannot make an uninformed decision, not to make it for them.

The second line of defence provides independent non-objection, not co-approval. Its role is to confirm that the first-line process was conducted with rigour, that the residual rating is defensible, and that proposed conditions are adequate to bring risk within appetite. Non-objection is a governance checkpoint, not shared ownership of the decision, and the distinction is not semantic: a second line that co-approves has taken on accountability for an outcome it does not control, which weakens rather than strengthens the first line's ownership of its own risk. One boundary on that position has to be stated, because a supervisory standard constrains it. For a large national bank the OCC's Guidelines Establishing Heightened Standards require independent risk management to "oversee the covered bank's risk-taking activities and assess risks and issues independent of front line units," and to take primary responsibility for the risk governance framework [12], so non-objection allocates the production decision without transferring that accountability: the second line remains answerable for the independence and quality of the challenge it performed, and withholding non-objection is the exercise of that accountability rather than an exception to it.

Legal reviews for regulatory and contractual compliance, confirming that the deployment creates no legal exposure the assessment has not captured or the owner has not adequately addressed.

The formal production gate requires all three: owner acceptance, second-line non-objection, and Legal clearance. The absence of any one is not a shortcut. It is a production blocker.

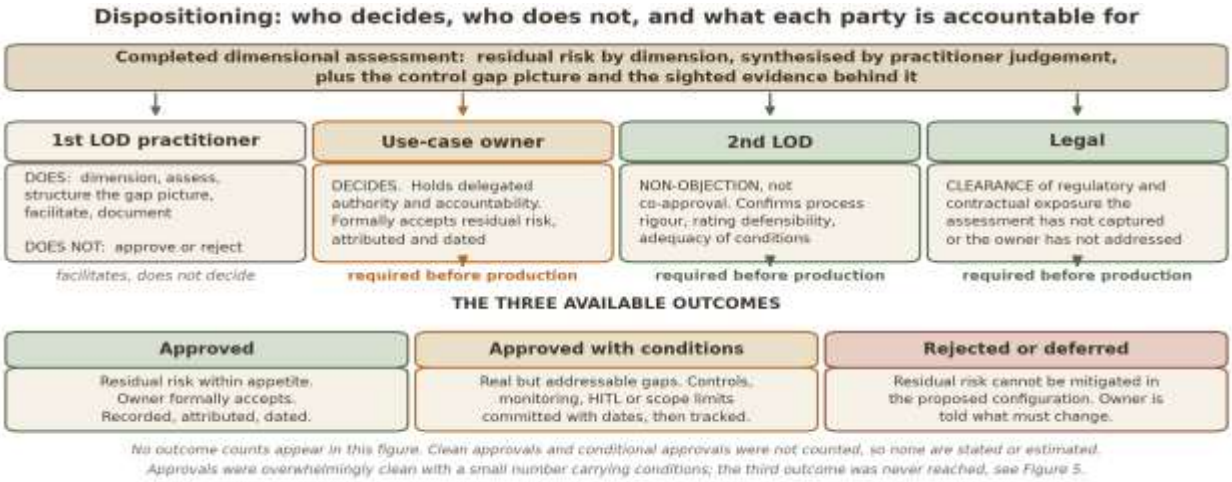


Figure 4. Decision rights at the production gate. The first line assesses, structures, and documents but does not decide; the owner decides and accepts; the second line does not co-approve; Legal clears a separate exposure. All three of the required roles must be satisfied before production.

11.3 What the decision rests on

A disposition without documented evidence is not a governance outcome. It is an opinion. The evidence base begins with the inherent risk assessment, a documented record of which risk scenarios were evaluated, which were assessed as materially applicable, and which were scoped out. That record is what makes the scoping judgement of Section 5 auditable rather than tacit: a later reviewer sees not only what was assessed but what was deliberately not assessed, and why. The remainder is the set of sighted artefacts specified per dimension in Table 2. The disposition record then adds the residual rating by dimension and in aggregate, the cross-dimensional dependencies between those ratings (Section 6.2), the control gaps identified, any conditions committed to with owners and dates, and the attributed acceptance, non-objection, and legal clearance.

12. Failure Scenarios

The absence of a structured risk gate does not produce a single catastrophic failure. It produces a slow accumulation of unresolved risk that surfaces in three predictable ways: incidents that were preventable, audit findings that were avoidable, and rework that upstream governance would have made unnecessary. The three scenarios below are realistic composites, constructed to illustrate failure patterns and the dimensions that catch them. They are not accounts of specific deployments, no institution is identified, and no cost comparison is asserted for any of them (Section 14). None is exotic: each is the predictable consequence of letting speed to

production override structured risk assessment, and in each the institution arrives at the supervisor's question of Section 1 with no answer prepared before the fact.

The model never scoped to its data risk. A machine learning model automating a credit-related decisioning workflow performs well in testing and goes to production without a data risk assessment: nobody asks where the training data comes from, whether it is representative, or whether it carries bias that will disadvantage certain customer segments. Six months in, audit flags that its outputs correlate with demographic proxies in ways that create fair lending exposure, and the model is pulled, retrained, re-validated, and resubmitted through an approval process that should have run first. Lineage validation, bias testing, and fair lending control checks are standard in a structured assessment (Table 2, Dimension 2) and surface this before production. That input-level data bias produces disparate credit outcomes is established [15], which is why the check attaches at the data layer.

The generative tool with no IAM control validation. A generative assistant summarising documents and answering policy questions is treated by its owner as a productivity tool rather than an AI system carrying technology risk, and goes live without a technology controls assessment. Nobody validates what the model can reach through its retrieval layer or whether that layer is scoped to the querying user's authorisation. Within weeks, users find it surfacing content from repositories cutting across organisational boundaries, visible through the AI interface but not directly. That is not a breach in the technical sense. It is a control failure, and it generates an incident, a technology risk finding, and an offline period while access controls are redesigned. IAM scoping, access boundary validation, and retrieval layer authorisation are standard technology controls (Section 9), and a structured assessment catches the misconfiguration first, on the evidence standard that matters here: a confirmed technical configuration, not a design document asserting the restriction.

The agentic system that operated beyond its intended scope. An agentic system automating a back-office workflow is framed as automation rather than AI and receives minimal scrutiny. Nobody asks what actions it can take autonomously, what the blast radius boundaries are, whether human-in-the-loop checkpoints sit at the right decision points, or what rollback looks like at scale. It encounters an edge case its training did not cover and, rather than failing safely and escalating, interpolates, executing a course of action plausible given its training and incorrect given the operational context; operating autonomously and at speed, the error propagates across hundreds of records before anyone notices. Remediation requires manual correction of every affected record, regulatory notification, and a retrospective concluding that the system should not have reached production without operational risk controls validation. Agentic systems carry autonomous action scope, human oversight checkpoints, failure mode behaviour, and rollback capability as load-bearing controls, which a structured assessment surfaces before deployment. The failure mode is the one the agentic safety literature identifies as characteristic of increasingly agentic systems, consequences propagating through actions rather than outputs [18]. Behavioural change after approval is a runtime problem rather than a gate problem [2].

13. Application and Observations

This section reports the operating record, narrowly about what is attributable to the author personally as against the organisational programme, and explicitly about what was not measured.

Table 3. Operating record. Personal scope and organisational programme scope are reported separately and must not be conflated.

Item	Recorded position
Author's personal scope	15 use cases led end-to-end through the full six-step methodology. Every observation in this section concerning the conduct of an assessment derives from these
Organisational programme context	Approximately 1,300 use cases governed through the process at organisation level, of which approximately 800 were generative or agentic, roughly 60 percent of the portfolio. This is the scale of the programme the methodology operated inside, not the author's personal caseload, and no claim in this paper rests on it
Period	Approximately early 2024 to mid-2026, continuous operational practice

Cycle time	**Organisational scope.** Four to six weeks per use case assessment
Approved clean	**Organisational scope.** The large majority. No exact count is recorded and none is estimated here
Approved with conditions	**Organisational scope.** A small number, the exception rather than the rule. No exact count is recorded and none is estimated here
Rejected or deferred	**Organisational scope.** None, across the approximately 1,300. See Section 13.1
Withdrawn, paused, or descoped during assessment	**Not captured at either scope,** and not reconstructable. See Section 13.1
Risks reaching production undetected	**Organisational scope.** No instances reported in which a material risk went undetected before production and was later found by audit or a supervisor. A reported absence rather than a measured finding, and the functions that would detect such a case partly overlap the function described here
Internal audit outcomes	**Organisational scope.** One audit finding, on AI system inventory completeness. See Section 13.2
Before-and-after comparison	Not measured at either scope. No instrumented baseline exists from before the methodology was in place

Every row carries its scope because in the source record the outcome figures were reported at organisation level and the author's figure separately. No outcome row is a distribution over the author's 15 use cases, and no claim in this paper rests on the organisational totals.

13.1 Zero rejections, and where the gate actually does its work

A production gate that has never rejected anything invites an obvious question: is it a gate at all? The record has to be read carefully, and hiding it would be worse than explaining it.

No use case was rejected or deferred anywhere in the organisational programme over the period, and none was among the 15 the author led. The reason offered is not that residual risk was waved through. It is that use-case owners resolved risk during the assessment rather than pushing to production with unresolved gaps. By the time a disposition decision was reached, the material gaps had been closed, which is why the outcome distribution sits between approved clean and approved with conditions rather than extending to rejection.

This is a finding about where in the process the gate does its work, consistent with the mechanism argued in Section 2. The rejection branch is the gate's last resort, not its operating mode. Its practical function is upstream: once owners understand that controls will be verified against evidence rather than accepted on assertion, and that residual risk will be attributed to them by name and date, the incentive is to resolve gaps during assessment. The gap picture produced in Step 2 of dispositioning is therefore the instrument doing most of the work. Two caveats attach. The mechanism is an interpretation of the observed outcome distribution, not a measured causal finding, because no counterfactual was run. And a rejection branch never exercised is a branch whose operation is untested, which Section 14 records as a limitation rather than a strength.

A third point is more damaging and is stated plainly rather than left for a reader to find. Withdrawal counts were not captured. Whether any use case was withdrawn, paused indefinitely, or descoped by its owner during assessment was not recorded at either scope and cannot be reconstructed now. Nor were the other indicators of upstream pressure: forced design changes, gaps identified per assessment, gap-closure times, or owner escalations. A gate reporting no rejections, no deferrals, and no withdrawals has no observable footprint in its own outcome distribution, so the effect claimed for it upstream cannot be demonstrated from that distribution alone, and this paper does not claim otherwise.

The record is therefore equally consistent with a second reading, named here rather than deflected: a gate that simply never says no. The reason for preferring the upstream reading is mechanical rather than statistical. Residual risk was attributed to a named owner on a dated record and controls were verified against sighted artefacts, so an owner arriving at the disposition step with unresolved gaps arrives at a record that names them for it, which changes the incentive before the step is reached. That is an argument about the mechanism, not

evidence about the outcome, and both readings are left standing because the record cannot close the gap between them. Section 14 lists what would.

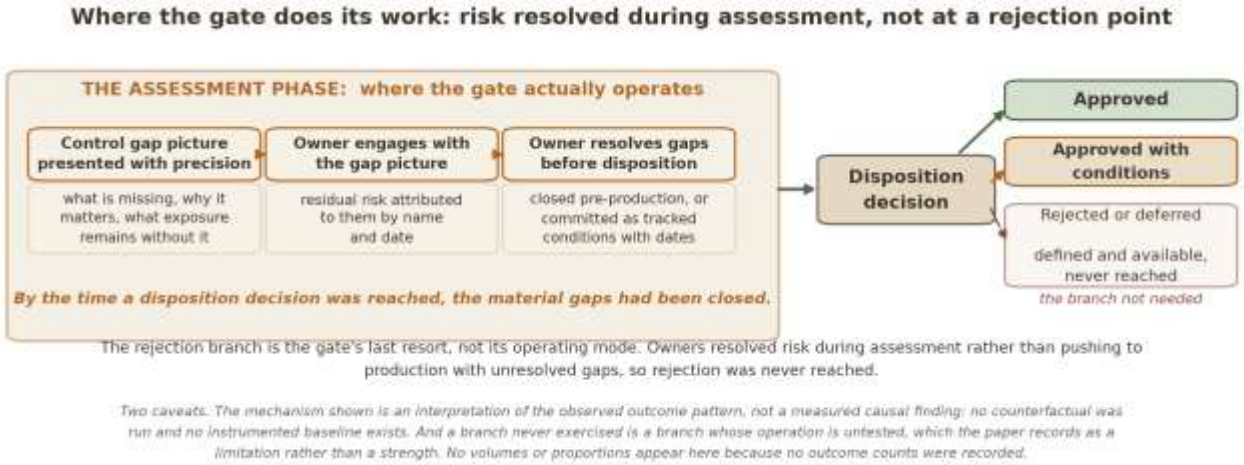


Figure 5. Where the gate does its work. Risk was resolved during assessment rather than at a rejection point, so the rejection branch, though defined and available, was never reached. This is an observation about the process, and the unexercised branch is recorded in Section 14 as a limitation.

13.2 The one audit finding, and what it was not about

One internal audit finding arose over the period. It concerned AI system inventory completeness, specifically the difficulty of maintaining visibility into all agents, Model Context Protocol servers, agentic skills, and third-party managed AI services.

The finding was not about the assessment methodology. It was about the predecessor problem, knowing what exists before you can govern it, which is the boundary condition on any pre-production gate. A gate governs what is brought to it and cannot govern what is never declared, and the completeness of the inventory feeding it is a separate control problem with separate mechanisms, registry enforcement in the development lifecycle [4] and discovery of unauthorised AI tool servers [5]. The methodology assumes a declared use case as its input, and the finding is the empirical reminder that the assumption is not free.

13.3 What was observed and what was not

Three observations are supportable from the record. First, the four-to-six-week cycle time is what running the full six steps to an evidenced disposition took in practice, reported as an operating characteristic rather than an improvement, because there is nothing to compare it against. Second, running the process consistently produces a risk inventory that surfaces patterns: which control domains are persistently weak, which risk dimensions are underweighted, and where the same gaps keep reappearing. That intelligence feeds risk appetite decisions and control investment prioritisation as a by-product of dimensioning consistently rather than a separate exercise. Pattern detection at portfolio scale is a property of the organisational programme rather than of a 15-case personal record, and is reported here as such. Third, the dimensional structure creates comparability across structurally different systems: a generative system and a classical machine learning system are very different things, but once both are assessed against the same framework their residual profiles sit on the same scale and can be governed in one forum in one language.

What was not measured must be equally clear (Section 14). Clean against conditional approvals were not counted and are not estimated, and neither were withdrawals, forced design changes, gaps per assessment, gap-closure times, or escalations.

14. Limitations

AI systems that do not fit the standard dimensions. The framework was designed around systems with a recognisable architecture: a model, a data source, a control environment, an operational workflow, and an identifiable owner. Most enterprise deployments fit that profile reasonably well. Some do not. The hardest cases blur the boundary between AI and conventional software: embedded AI features inside vendor software-as-a-

service platforms, AI capabilities surfaced through productivity tools, and AI components woven into legacy systems without clear demarcation. There the intake step itself is contested, because there may be no identifiable owner of the AI capability as distinct from the owner of the platform hosting it, and the dimensioning inherits that ambiguity.

The dimensional set is not exhaustively specified here. The framework comprises five dimensions and this paper develops three (Section 6.1); the remaining two are part of the methodology as operated but are not specified here, so Sections 7 to 9 are not the complete set. Within the three developed, Dimension 3 is presented at the level of its control domains, their AI-specific recalibration, and the verification standard, rather than an exhaustive per-domain evidence specification. That scope also bounds the first contribution claim: Section 6.2 works three interaction pairs where a five-dimension framework contains ten, so the integration layer is demonstrated across part of the set rather than all of it.

Single institution, and the limits of the personal record. The methodology was operated inside one large financial institution, against its four risk frameworks and its AI governance standard. Transferability to institutions with a different control taxonomy, three-lines structure, or regulatory perimeter is untested. The author's personal record is 15 use cases led end-to-end; the approximately 1,300-use-case organisational programme is the context the methodology operated inside, not a personal sample.

No baseline, and therefore no comparative evidence. No instrumented baseline exists from before the methodology was in place. Every comparative proposition here is an argument rather than a finding, including the claim in Section 2 that a functioning gate accelerates adoption by moving issue discovery upstream, and no cost claim is made about the composites in Section 12.

The rejection branch is untested, and the gate has no observable footprint. No use case was rejected or deferred (Section 13.1), so the reject-and-resubmit path, including how a rejected owner's escalation is handled and how a resubmission is re-dimensioned, has not been exercised: its design is specified, its operation is not evidenced. The more consequential absence is on the other side of the same record. Withdrawals, pauses, and descopings during assessment were not captured, and neither were forced design changes, gaps per assessment, gap-closure times, or escalation counts. Each is a by-product a governance process generates, none was recorded, and the result is that the upstream reading of the zero cannot be separated from the sceptical one. Recording any one of them would change that, and it is the first thing this methodology should instrument. Clean against conditional approvals were also not counted, which precludes any analysis of which control domains most often drive conditions.

Residual risk synthesis is a documented judgement, not a reproducible calculation. The methodology deliberately rejects arithmetic aggregation in favour of a practitioner judgement about dimensional interaction (Sections 6.2, 11.1). The cost is that reproducibility across practitioners is asserted rather than demonstrated: no inter-rater study was conducted, even though the documentation such a study would need is produced by the process itself. Appendix A states what an institution needs to instantiate the method, and which instruments are not released here.

The evidence stops at production clearance, and inventory completeness is assumed. The methodology terminates at the production decision. It says nothing about whether controls continued to operate afterwards, whether conditions were closed on their committed dates, or how systems behaved once live; that layer is addressed in the author's separate work [2]. At the other end, a gate governs only what is declared to it, and as Section 13.2 records the one audit finding concerned inventory completeness. Mechanisms for ensuring completeness of the declaration sit outside this methodology [4], [5].

15. Conclusion

Financial institutions do not need a new framework to govern AI. They have model risk management, technology control, data risk, and operational risk frameworks, and increasingly an AI standard layered above

them. What has been missing is the mechanism that converts those frameworks into a production decision for a specific use case, with a defensible record of what was assessed, what was verified against evidence, what exposure remained, and who accepted it.

Risk dimensioning and dispositioning is that mechanism. Dimensioning decomposes the use case's risk across distinct domains and asks what exposure remains once controls are accounted for. Dispositioning closes the assessment into a documented decision in which the use-case owner accepts residual risk, the second line provides non-objection rather than co-approval, and Legal clears regulatory and contractual exposure, all three required before production. What makes both halves real rather than documentary is that the practitioner sights the artefact instead of accepting the assertion. Three contributions are claimed and no more: the integration layer treating the dimensions as interacting rather than parallel and recording the dependencies between them, the formalisation of the scoping judgement, and the allocation of decision rights at the production gate, which the three-outcome close carries rather than constitutes.

Three boundaries close the paper honestly. The framework comprises five risk dimensions and this paper develops three. Because no instrumented baseline exists, nothing here claims the methodology is faster, cheaper, or better than what preceded it. And because withdrawals, forced design changes, and gap-closure times were never captured, the gate's effect cannot be demonstrated from its outcome distribution at all. What can be said is narrower, and it describes the organisational programme rather than the author's 15-case personal scope, which Table 3 reports separately: AI systems reached production with a documented, auditable risk record; the material gaps appear on that record as resolved during assessment rather than at the gate, which is an interpretation and not a measured causal finding; and the one audit finding that arose was about knowing what exists rather than how what was known had been assessed. That is the floor a regulated institution needs. It is not the ceiling.

Appendix A. What an Institution Needs to Instantiate the Methodology

Section 14 records that reproducibility across practitioners is asserted rather than demonstrated. This appendix does not repair that. It states what an institution has to have in place before the six steps of Section 5 can be run at all.

Table 4. Prerequisites for instantiating the methodology.

Prerequisite	What it must supply
Named risk frameworks with enumerated risk scenarios	Step 1 scopes which scenarios apply, so they must already be enumerated to scope against. The methodology orchestrates existing frameworks rather than replacing them (Sections 3.2, 5)
A declared-use-case intake, and an identified owner	A gate governs only what is brought to it, so without a mechanism surfacing undeclared AI workloads the inventory gap of Section 13.2 remains; those mechanisms sit outside this methodology [4], [5]. The disposition belongs to an owner with delegated authority (Section 11.2)
Evidence standards stated in advance, and an auditable record store	The specific artefact sighted for each control asserted, per dimension as in Table 2; and a store in which the disposition record is the one an auditor or supervisor would ask for, not a parallel document produced for review (Section 10)
A residual rating scale, and a synthesis convention rather than a formula	A defined scale, typically low, moderate, high, critical, each level carrying a defined consequence at the gate (Section 11.1), institution-specific and not reproduced here. No arithmetic aggregation, and instead the requirement that the residual record state cross-dimensional dependencies explicitly, in the form Section 6.2 works
Three separable decision roles, and instrumentation this instantiation lacked	Owner acceptance, second-line non-objection, and Legal clearance, all three required, with the first line facilitating but not deciding; where the second line co-approves instead, the allocation claimed in Section 3.2 is not what is being instantiated. Capture withdrawals, forced design changes, gaps per assessment, gap-closure times, and escalations from the first assessment: none was captured here (Sections 13.1, 14)

Three instruments are deliberately not reproduced, and their absence is the practical limit on reproducibility: the intake and scoping worksheet, the per-level definitions of the rating scale, and a worked end-to-end assessment are institution-specific artefacts of the deployment described. A second practitioner given this appendix can stand the process up; whether they would reach the same residual ratings on the same use case is the open question Section 14 states.

References

[1] S. K. Anuguthala, "Enterprise Agentic AI Lifecycle Governance: A Control-Driven Framework from Design to Decommissioning," *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, vol. 7, no. 2, pp. 9-16, 2026. doi: 10.63282/3050-9262.ijaidsml-v7i2p103.

[2] S. K. Anuguthala, "Guard: A Governance-Anchored Framework for Runtime Monitoring and Incident Response in Enterprise Agentic AI Systems," *International Journal of AI, BigData, Computational and Management Studies*, vol. 7, no. 2, pp. 370-379, 2026. doi: 10.63282/3050-9416.ijaibdcms-v7i2p147.

[3] S. K. Anuguthala and H. Namani, "Enterprise AI Supply Chain Security Governance: A Framework for Secure Open-Source AI Model Adoption Using an Artificial Intelligence Bill of Materials (AI-BOM)," *International Journal for Multidisciplinary Research*, vol. 8, no. 3, 2026. doi: 10.36948/ijfmr.2026.v08i03.78068.

[4] S. K. Anuguthala, "Preventing Shadow AI Workloads at Enterprise Level: A Novel Integration of AI Governance into Software Development Lifecycle (SDLC)," in *World Forum on Cybersecurity, Machine Learning, and Intelligent Systems*, pp. 61-75, 2026. doi: 10.63282/3117-5481/wfcmls26-107.

[5] S. K. Anuguthala, "Shadow MCP: Detecting and Governing Unauthorized AI Tool Servers in Enterprise Environments," *Journal of Advances in Developmental Research*, vol. 17, no. 1, 2026. doi: 10.71097/ijaidr.v17.i1.2034.

[6] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, *Supervisory Guidance on Model Risk Management*, 21 pp., dated Apr. 4, 2011; the attachment to Federal Reserve SR letter 11-7, and issued jointly as OCC Bulletin 2011-12, "Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management," Apr. 4, 2011. [Online]. Available: https://www.federalreserve.gov/boarddocs/srletters/2011/sr1107a1.pdf (dated 2011 document; accessed Aug. 10, 2026). All quotations of the 2011 guidance are from this document; the Federal Reserve's undated guidance landing page now serves the 2026 revision at the same address.

[7] Board of Governors of the Federal Reserve System, "Revised Guidance on Model Risk Management," SR letter 26-2, Apr. 17, 2026, superseding SR letter 11-7 (Apr. 4, 2011) and SR letter 21-8, "Interagency Statement on Model Risk Management for Bank Systems Supporting Bank Secrecy Act/Anti-Money Laundering Compliance" (Apr. 9, 2021); and OCC Bulletin 2026-13, "Model Risk Management: Revised Guidance," Apr. 17, 2026, rescinding OCC Bulletin 2011-12. [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/SR2602.pdf (dated 2026 document; accessed Aug. 10, 2026).

[8] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Jan. 2023, doi: 10.6028/NIST.AI.100-1; and *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1, Jul. 2024, doi: 10.6028/NIST.AI.600-1.

[9] Federal fair lending authorities: Equal Credit Opportunity Act, 15 U.S.C. sec. 1691 et seq., and Regulation B, 12 CFR pt. 1002 (adverse action reasons at 12 CFR 1002.9(b)(2), 12 CFR ch. X, 1-1-25 ed.); Fair Housing Act, 42 U.S.C. sec. 3601 et seq.; and Consumer Financial Protection Bureau, "Equal Credit Opportunity Act Baseline Review Modules," in *CFPB ECOA Examination Procedures Baseline Review*, Apr. 2019. [Online]. Available: https://www.govinfo.gov/content/pkg/CFR-2025-title12-vol8/pdf/CFR-2025-title12-vol8-sec1002-9.pdf ; https://files.consumerfinance.gov/f/documents/cfpb_supervision-and-examination-manual_ecoa-baseline-exam-procedures_2019-04.pdf (accessed Aug. 10, 2026).

[10] Interagency Guidelines Establishing Information Security Standards, under GLBA sec. 501(b) and FDI Act sec. 39; codified at 12 CFR pt. 30 app. B, pt. 208 app. D-2, pt. 225 app. F, pt. 364 app. B. Quoted from 12 CFR pt. 30 app. B secs. II.A, III.C.1, 12 CFR ch. I, 1-1-25 ed. [Online]. Available: https://www.govinfo.gov/content/pkg/CFR-2025-title12-vol1/pdf/CFR-2025-title12-vol1-part30-appB.pdf (accessed Aug. 10, 2026).

[11] Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency, *Interagency Guidance on Third-Party Relationships: Risk Management*, 23 pp., dated Jun. 7, 2023, published in the *Federal Register* Jun. 9, 2023; transmitted as Federal Reserve SR letter 23-4 and OCC Bulletin 2023-17. [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/SR2304.pdf (dated 2023 document; accessed Aug. 10, 2026).

[12] OCC Guidelines Establishing Heightened Standards for Certain Large Insured National Banks, Insured Federal Savings Associations, and Insured Federal Branches, 12 CFR pt. 30 app. D secs. II.A, II.C.2, 12 CFR ch. I, 1-1-25 ed. [Online]. Available: https://www.govinfo.gov/content/pkg/CFR-2025-title12-vol1/pdf/CFR-2025-title12-vol1-part30-appD.pdf (accessed Aug. 10, 2026).

[13] Data protection: Regulation (EU) 2016/679 (General Data Protection Regulation), arts. 5(1)(b), 5(1)(e), 17, *Official Journal of the European Union*, L 119, pp. 1-88, May 4, 2016; and California Consumer Privacy Act, Cal. Civ. Code sec. 1798.105 (deletion right, as amended by Proposition 24, Nov. 3, 2020, operative Jan. 1, 2023). [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj ; https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.105 (accessed Aug. 10, 2026).

[14] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1-37, Mar. 2014. doi: 10.1145/2523813.

[15] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *The Journal of Finance*, vol. 77, no. 1, pp. 5-47, 2022. doi: 10.1111/jofi.13090.

[16] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, et al., "Red Teaming Language Models with Language Models," in *Proc. 2022 Conf. on Empirical Methods in Natural Language Processing*, pp. 3419-3448, 2022. doi: 10.18653/v1/2022.emnlp-main.225.

[17] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," in *Proc. 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79-90, 2023. doi: 10.1145/3605764.3623985.

[18] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, et al., "Harms from Increasingly Agentic Algorithmic Systems," in *Proc. 2023 ACM Conf. on Fairness, Accountability, and Transparency*, pp. 651-666, 2023. doi: 10.1145/3593013.3594033.

[19] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, et al., "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in *Proc. 2020 Conf. on Fairness, Accountability, and Transparency*, pp. 33-44, 2020. doi: 10.1145/3351095.3372873.

[20] H. Davies and M. Zhivitskaya, "Three Lines of Defence: A Robust Organising Framework, or Just Lines in the Sand?," *Global Policy*, vol. 9, no. S1, pp. 34-42, Jun. 2018. doi: 10.1111/1758-5899.12568.