International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 1743-1756
https://doi.org/10.70917/ijcisim-2026-5726
Article

A Structured Evaluation Framework for AI Tool Adoption in Regulated Financial Decisioning: Quantifying Model Risk Before Production Use

Sushmita Vinod Naik

Independent Researcher,India
Email: sushmitavnaik@gmail.com

Abstract: Financial institutions increasingly place artificial intelligence tools into decisions carrying regulatory consequence, including collateral valuation, credit assessment, and tenant selection. Model risk guidance addresses tools already in use, and the April 2026 revised interagency guidance expressly excludes generative and agentic systems while narrowing its stated relevance to the largest banking organisations. A gap therefore exists in time and in coverage: legal consequence attaches at first use, while governance engages afterwards, and institutions outside supervisory scope remain fully subject to the substantive consumer protection statutes. This paper proposes a pre-adoption framework that characterises the decision a tool will influence, scores inherent risk across seven weighted dimensions, applies capped control credit to yield a residual exposure score, and maps that score to a documented adoption gate. Two veto rules and one condition rule override the arithmetic where a legal obligation is categorical rather than tradeable, which is the framework's principal design contribution: weighted additive scoring is compensatory by construction, and some obligations are not compensable. The framework is proposed, not validated. It has never been applied to a real tool, no institution has run a tool through it, no adoption decision anywhere has been informed by it, and no claim is made that using it produces better decisions than the practice it would replace. That is a hypothesis the paper sets out to be tested. The two worked examples are synthetic constructions with invented inputs, and the paper reports no results.
Keywords: model risk management, artificial intelligence governance, real estate finance, automated valuation models, fair lending, technology adoption decisions, pre-adoption assessment, weighted scoring, non-compensatory decision rules, adverse action

1. INTRODUCTION

1.1 A gap in time, not in rigour

Model risk frameworks are mature and, within their scope, effective. Their scope is tools in use. Validation, monitoring, and periodic review all presuppose something to validate, monitor, and review.

Legal consequence does not wait for that. The obligation to explain an adverse action exists from the first decision a tool influences, not from the point at which an institution's validation cycle happens to conclude. Any process engaging after adoption is structurally late for the obligations that matter most, and being rigorous does not fix being late.

Figure 1. The timing gap. Legal consequence attaches at first use; governance engages afterwards.

There is a coverage gap alongside the timing gap. Supervisory guidance is written for banking organisations and the April 2026 revision states it is most relevant to those above thirty billion dollars in assets, while expressly excluding generative and agentic systems. Non-depository lenders, debt funds, servicers, sponsors, and property managers remain fully subject to the substantive consumer protection statutes while sitting outside that guidance. The revision made this more pressing rather than less.

1.2 Status of this work

This is a framework proposal. It has not been applied.

The manuscript was drafted in July 2026. No institution has run a tool through the framework. No adoption decision has been informed by it. It has not been validated against outcomes, and no claim is made that using it produces better decisions than existing practice. That is the hypothesis the paper sets out to be tested, not a finding it reports.

Three categories of content should be kept apart, because conflating them is how a design paper becomes a false empirical claim.

Design. The four-stage structure, the seven risk dimensions, the control credit catalogue, the residual exposure formula, the four adoption bands, the two veto rules, and the condition rule are constructs the author designed. The specific numbers, meaning the dimension weights, the control credits, the band thresholds, and the cap on control credit, are stipulated starting values. They were chosen to reflect the author's reading of where regulatory consequence concentrates in real estate finance. They were not elicited from a panel, not estimated from data, and not calibrated against outcomes. A different person with comparable experience might reasonably choose different values, and in a case near a band boundary that would change the outcome.

Practitioner experience. One element has a different basis. Dimension D7, integration fragility, and the condition rule attached to it derive from the author's experience building and operating document extraction tooling, specifically the failure mode where a tool continues producing well-formed, confident output after an upstream change has invalidated it. That is an experiential basis for a design choice. It is not an application of the framework and is not a case study, pilot, or validation.

Illustration. The two worked examples in Section 7 are synthetic constructions. The lender does not exist and the tools do not exist. Every dimension score is a number assigned to make the mechanics visible. The resulting figures are arithmetic outputs of invented inputs. The arithmetic is correct; correct arithmetic on invented numbers produces an illustration, not a measurement.

1.3 Contribution

1. A decision-first unit of analysis, where the same tool can be appropriate in one decision and inadmissible in another.

2. Pre-adoption timing, engaging before the point at which existing frameworks do.

3. Institution-agnostic applicability, usable by entities outside supervisory guidance written for large banks.

4. Non-compensatory overrides, which is the principal design contribution.

5. A distinction between risks resolvable at the gate and risks that are not, expressed through condition rules.

What is not new, and the paper should not pretend otherwise: the dimensions themselves are largely familiar from existing risk literature. The contribution is in the assembly, the timing, and the override logic, not in discovering that explainability or data provenance matter.

2. How Adoption Decisions Actually Get Made

In observation, the adoption of an AI tool is frequently not made as a risk decision at all. It is made as a procurement decision: someone evaluates vendors, negotiates terms, and signs a contract, and the risk question, if it arises, arises as a compliance checkbox within purchasing rather than as substantive analysis.

Alternatively it happens through pilot drift. A team runs a trial. The trial goes reasonably well. Nobody formally ends it. The tool acquires users, then acquires reliance, and at some point the output stops being advisory and becomes the basis of the decision. There is no moment at which anyone decided that transition should happen. Governance machinery typically engages after this point, which means the institution is validating a tool it has already committed to rather than deciding whether to commit. Those are different exercises with different incentives, and the second is much harder to fail.

Three independent reasons ad-hoc evaluation is insufficient here.

Timing. Legal consequence attaches at first use. Any process engaging after adoption is structurally too late.

Absence of a record. An ad-hoc evaluation produces a decision but not an artifact. When an examiner, counsel, or a counterparty in diligence asks why this tool was admitted to this decision and another was not, there is nothing to produce. The reasoning existed in conversation and the people have often moved on.

Invisible weighting. Unstructured judgement has the same weighting problem as structured scoring, without the transparency. Someone deciding informally that a tool is acceptable is weighting considerations against each other; they are simply not writing down how. The assumptions are still present, merely not contestable. This is the most under-appreciated argument for structure, and it should be stated as the modest claim it is: structure does not make the judgment correct. It makes it inspectable.

3. Failure Mechanisms

The following are mechanisms drawn from the regulatory record and published research, not incidents the author witnessed. No war story is offered because none is available, and four well-understood mechanisms are more useful than one vivid anecdote that cannot be stood behind.

The tool becomes authoritative without anyone deciding it should. This is the most characteristic case, because there is no moment of failure to point at. A valuation tool is introduced as a cross-check. Analysts find it faster than assembling comparables by hand. Within months the tool's number is the number in the file. The human review that justified its introduction has become a formality. Nobody decided the tool should be authoritative; the status changed through use. The consequence is that no assessment was ever triggered, because from the institution's point of view no decision was made. If the transactions are covered by the interagency valuation rule, the institution has by then acquired duties covering confidence in estimates, protection against data manipulation, conflicts of interest, random

sample testing, and nondiscrimination compliance, which it has never scoped, staffed, or documented. It is not that it accepted those obligations and discharged them poorly. It never knew it had them.

Discovering an explainability problem after the fact. A model demonstrably improves accuracy and is adopted on that basis. Later an applicant is declined and the institution must state the specific principal reasons. The vendor treats the weighting logic as proprietary, or the tool produces reason codes that are post-hoc approximations nobody has validated, or the reasons are technically true and meaningless to the applicant. Supervisory guidance has been direct that adverse action requirements apply regardless of technology, that an inability to understand one's own methods is not a defence, and that reliance on post-hoc explanation requires validating the accuracy of those approximations. What makes this costly is that exposure attaches per adverse action and is retroactive. It is not a forward problem fixable by process change; every decision the tool has already influenced is potentially affected.

Aggregate performance concealing distributional harm. A tool performs better on average and is adopted for that reason. Average performance is the natural thing to measure and often the only thing a vendor reports. Published research on United States mortgage data has found that machine learning default models produce distributional effects across borrower groups, with some groups disproportionately less likely to benefit, attributed primarily to the greater functional flexibility of these models rather than to proxy triangulation. The effect therefore arises from how the models work rather than from anyone supplying a prohibited variable. An institution measuring aggregate accuracy would see improvement and detect nothing.

Silent degradation. An extraction tool works correctly for a year. A third party then revises a document template: a contractor changes an invoice layout, a title company alters a form, a subcontractor sends scans rather than native files. One field stops populating. Every other field extracts normally. The output is well formed and looks exactly like correct output. The reviewer checks the values present; nothing prompts them to notice a value that is absent. This is the failure mode that motivated D7, because it is invisible in a way the others are not: the tool did not error, the output did not look wrong, and human review is structurally incapable of catching it. The reviewer and the tool share the same blind spot.

Why these four and not a longer list. The four mechanisms above were selected because each is documented in the regulatory record or published research rather than inferred, and because each fails in a way existing governance does not catch. Two of them share a property worth naming: in both the authority-through-use case and the silent degradation case, nothing errors, nobody is notified, and the institution's own controls report normal operation. Failure modes with that shape are systematically under-weighted in risk frameworks, because frameworks are generally built by reasoning backwards from incidents, and these produce no incident until they produce a large one.

On cost. No cost figure is given and none should be added. Costs vary enormously by institution size, transaction volume, and enforcement posture, and any figure stated here would be invented. A costing dimension, if needed, should come from published enforcement actions with citations and be presented as an illustrative range rather than an expected loss.

4. The Seven Dimensions

Model risk has two conventional sources: error in the model itself, and misuse of output that is otherwise sound. In regulated real estate finance a third element attaches. There is a legal consequence to either failure that is separate from the financial consequence and frequently larger. An erroneous valuation costs money. An erroneous valuation in a covered transaction that the institution cannot explain, or cannot show it tested for discriminatory effect, costs something else entirely, and that second cost does not scale with the size of the error.

Each dimension is scored 1 to 5, with 5 always the worse condition. Uniform polarity matters more than it sounds: mixed polarity is where scoring frameworks accumulate errors, because assessors lose track of which direction is bad. Anchors are defined at 1, 3, and 5; scores of 2 and 4 are intermediate and require written justification.

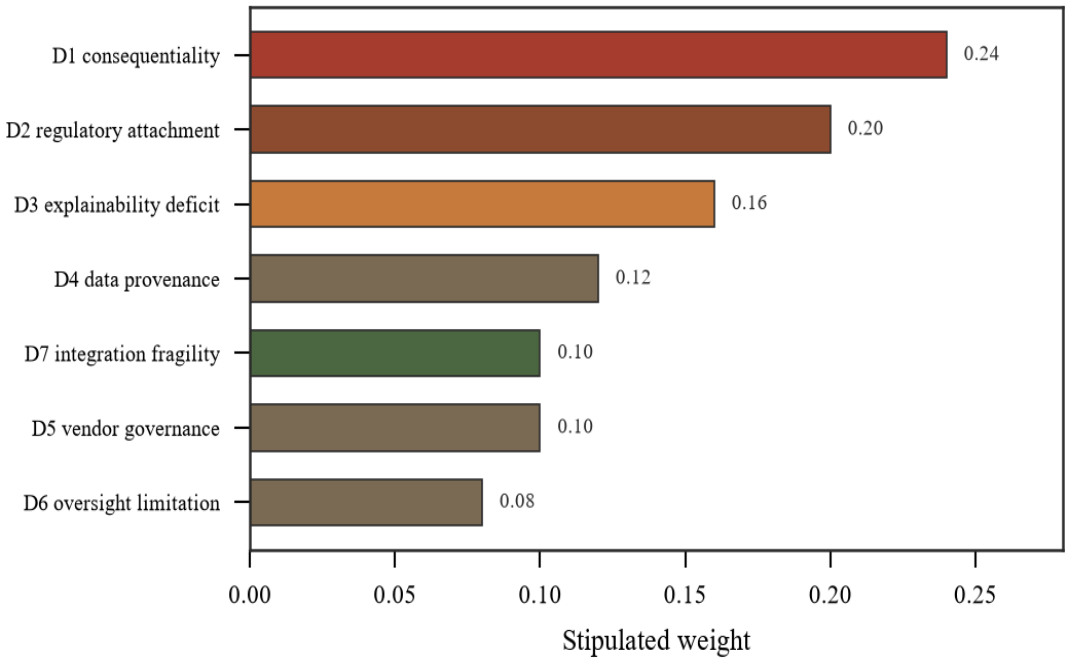

Figure 2. The stipulated dimension weights. These are starting values chosen by the author, not elicited from a panel or calibrated against outcomes.

Table 1. The seven dimensions with weights and anchors

Dimension	**Weight**	**Measures**	**Anchor 1**	**Anchor 3**	**Anchor 5**
D1 Decision consequentiality	0.24	Harm to the affected party if the tool is wrong	Internal or administrative; error readily corrected	Material financial consequence to an institutional counterparty; recoverable	Irreversible material harm to a consumer's housing, credit, or tenancy position
D2 Regulatory attachment	0.20	Density of legal obligation on the decision	No sector-specific obligation	Contractual, recordkeeping, or internal control obligations	Covered transaction with simultaneous adverse action and nondiscrimination exposure
D3 Explainability deficit	0.16	Gap between required and available reason-level explanation	Deterministic or directly traceable to source evidence	Opaque but each element independently verifiable against source	No validated reason-level explanation obtainable from tool or vendor
D4 Data provenance risk	0.12	Uncertainty about origin, coverage, local representativeness	Institution's own records with known coverage	Third-party data, disclosed sources, unverified local representativeness	Undisclosed provenance, no evidence of coverage in relevant submarkets

D5 Vendor governance risk	0.10	Ability to obtain what is needed to own the risk	Built in-house, or full documentation and testing rights	Standard assurance reporting, limited documentation, change notice by release note	Vendor refuses documentation or audit rights, model updated without notice
D6 Oversight limitation	0.08	Feasibility of meaningful review and of reversal	Every output reviewed before consequence attaches	Sample review feasible, reversible within a defined window	Review infeasible at volume, decision effectively final on output
D7 Integration fragility	0.10	Silent degradation after upstream change, and detectability	Inputs and interfaces institution-controlled under its own release process	Some inputs third-party controlled, change periodic and usually announced	Formats or interfaces change without notice, no surveillance of output distribution

Weights sum to 1.00.

Two dimensions warrant comment.

D3 is framed as a deficit deliberately. What matters is not whether the tool is interpretable in the abstract but whether it can produce what the law requires in this specific decision.

D7 is the dimension least represented in existing frameworks and the one a reader should most notice. It is genuinely distinct from D5 and D6, and the distinction is easy to miss. D5 asks whether the vendor can be inspected. D6 asks whether a human reviews the output. An institution can hold complete audit rights and review every single output and still be unable to see that a field silently stopped populating three weeks ago. What is failing is not accuracy and not oversight. It is detectability.

On the weights. The ordering of D1 above D2 above D3 above D5 reflects the author's judgement about where consequence concentrates in this sector. The placement of D4, D6, and D7 is less firmly established and would be revised on good argument.

The ordering attracting most scrutiny will be D1 above D2, since the paper's motivation is regulatory and the weighting says harm matters more than regulatory exposure. The reasoning is that regulation is a proxy for harm: the statutes exist because of the harms they prevent, so harm is the more fundamental quantity and regulatory attachment is a strong but imperfect indicator of it.

5. The Four Stages

Stage 1: characterise the decision. The assessor writes one sentence stating what decision the tool will influence and in what sense, then records four attributes: whose interests are affected and whether they are a consumer; whether an adverse action is possible; whether the decision is reversible before consequence attaches; and whether the transaction is covered by the interagency valuation rule.

This stage does most of the framework's work and should resist compression, because it forces a distinction vendor materials systematically obscure. A tool described as a valuation assistant might supply the number a credit decision turns on, or a preliminary estimate a licensed appraiser subsequently supersedes. Those are entirely different decisions with different regulatory attachment, and the difference is not a property of the software. It cannot be read off a datasheet. It has to be stated by someone who knows how the tool will actually be used.

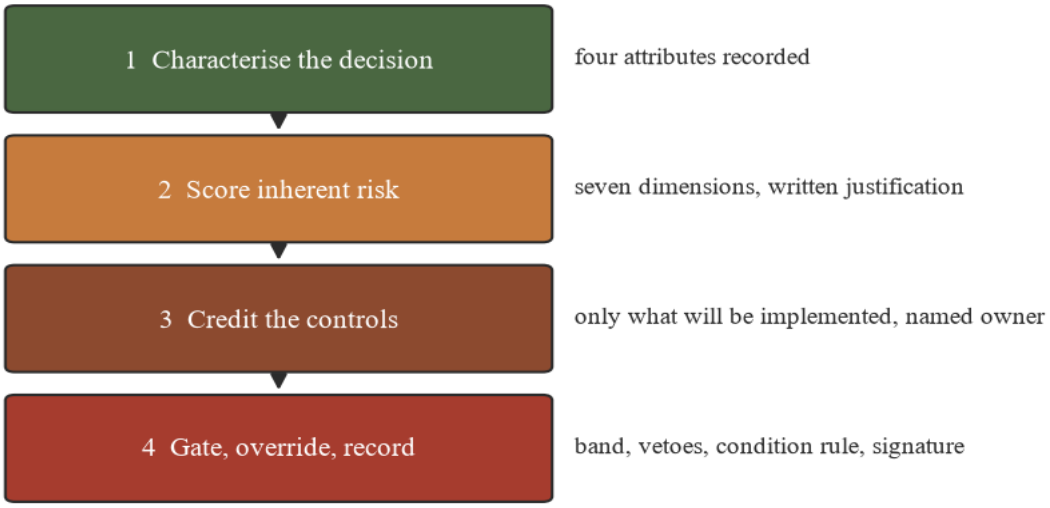

Figure 3. The four stages. Stage 1 does most of the work, because the same tool carries different risk in different decisions.

Stage 2: score inherent risk. All seven dimensions are scored against the anchors with written justification. The inherent model risk score is the weighted mean rescaled to a 0 to 100 range. The rescaling is presentational, allowing residual exposure and thresholds to be discussed on a familiar scale.

The four attributes recorded at Stage 1 are not descriptive metadata. Each maps directly to a dimension anchor or an override condition: whether a consumer is affected bears on D1, whether an adverse action is possible bears on D2 and triggers V1, reversibility bears on D6, and coverage by the valuation rule bears on D2 and V2. Recording them first means the scoring in Stage 2 is constrained by facts about the decision rather than by impressions of the technology.

Stage 3: credit the controls. Controls are credited where the institution will actually implement them, as distinct from controls it could in principle implement. That distinction is the entire integrity of the stage, and every control receives a named owner in the record, which is what makes the commitment real rather than aspirational.

Table 2. Control credit catalogue.

Control	**Failure mode addressed**	**Maximum credit**
Human review of every output before the decision	Undetected individual-case error	0.15
Confidence threshold with fallback to conventional method	Silent extrapolation beyond competence	0.12
Independent benchmark testing against a reference method	Systematic bias in output	0.12
Submarket-level nondiscrimination testing with remediation path	Disparate impact across protected classes	0.12
Tool cannot execute; output advisory to a named person	Automation of consequence without accountability	0.12

Random sample review at a defined rate, documented	Drift and undetected degradation	0.10
Field-level confidence flagging with exception routing	Uniform treatment of unequal-quality outputs	0.10
Continuous output-distribution and null-rate surveillance with alerting	Silent degradation after upstream change	0.10
Contractual documentation, testing rights, change notice	Vendor opacity and silent model updates	0.08

Credits are additive and capped at 0.60. Residual exposure is inherent risk multiplied by one minus the total credit.

Stage 4: gate, overrides, and record. Residual exposure maps to a band, the overrides are tested, and the record is written.

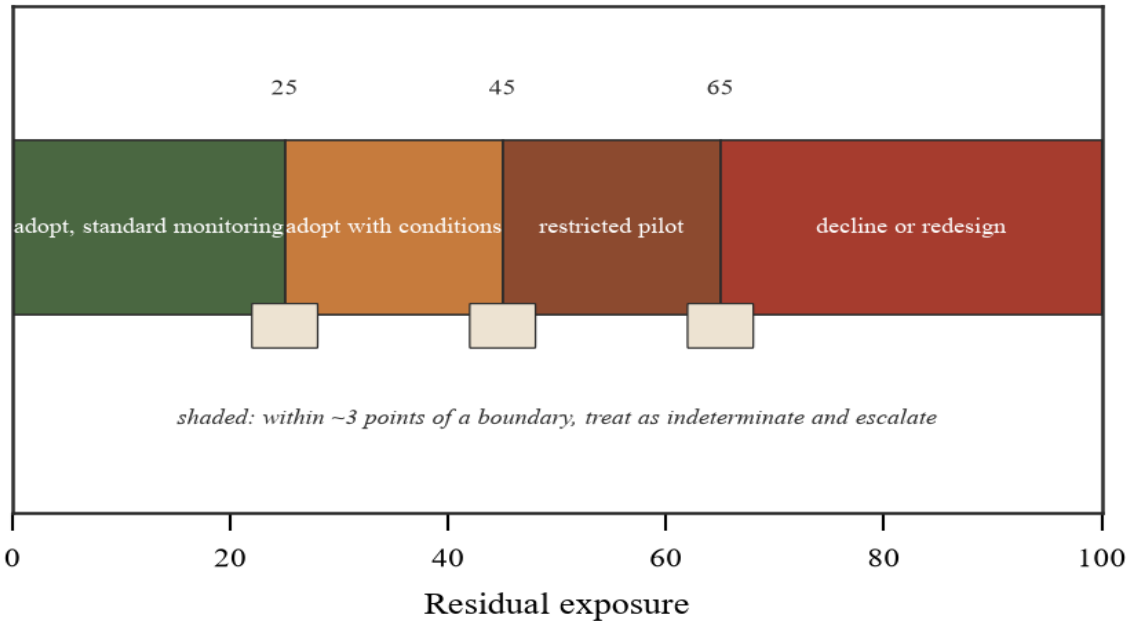

Figure 4. The four adoption bands and the near-boundary convention.

Table 3. Bands and override rules.

Trigger	**Condition**	**Effect**
Band 1	Residual exposure at or below 25	Adopt with standard monitoring
Band 2	Above 25 to 45	Adopt with conditions specified in the record
Band 3	Above 45 to 65	Restricted pilot; output may inform the decision but not constitute its basis
Band 4	Above 65	Decline, or return for redesign
V1	D2 equals 5 and D3 is 4 or higher	Cap outcome at restricted pilot regardless of residual exposure

V2	D5 equals 5	No authoritative use in any covered transaction
C1	D7 is 4 or higher	Continuous surveillance mandatory in every band; breach added to re-review triggers
Near-boundary	Residual exposure within roughly 3 points of a band edge	Treat as indeterminate; escalate rather than resolve arithmetically
Control cap	Total credit would exceed 0.60	Cap at 0.60

6. Why the Arithmetic Must Be Overridden

Two design choices deserve explicit defence, because they are what a reviewer will probe.

Why cap control credit. Controls mitigate model risk; they do not eliminate it. Without a cap, an institution could document its way to an arbitrarily low residual score without changing anything about the tool or the decision, simply by committing to enough controls on paper. The cap forces at least forty percent of inherent risk to remain visible in the record. Better to be accused of crudeness here than of being gameable.

Why multiplicative rather than subtractive. Controls reduce risk proportionally rather than absolutely. A high-risk tool with excellent controls should still carry more residual risk than a low-risk tool with merely adequate ones, and subtraction does not preserve that.

The overrides are the heart of the framework. Weighted additive scoring is compensatory by construction: strength anywhere offsets weakness anywhere. That is usually a feature, since it is what lets a framework express trade-offs. But some obligations are not tradeable, and where that is true the arithmetic must be overridden rather than adjusted. Tuning the weights cannot fix this, because the problem is not that a dimension is underweighted. The problem is that the dimension is not on a scale with the others at all.

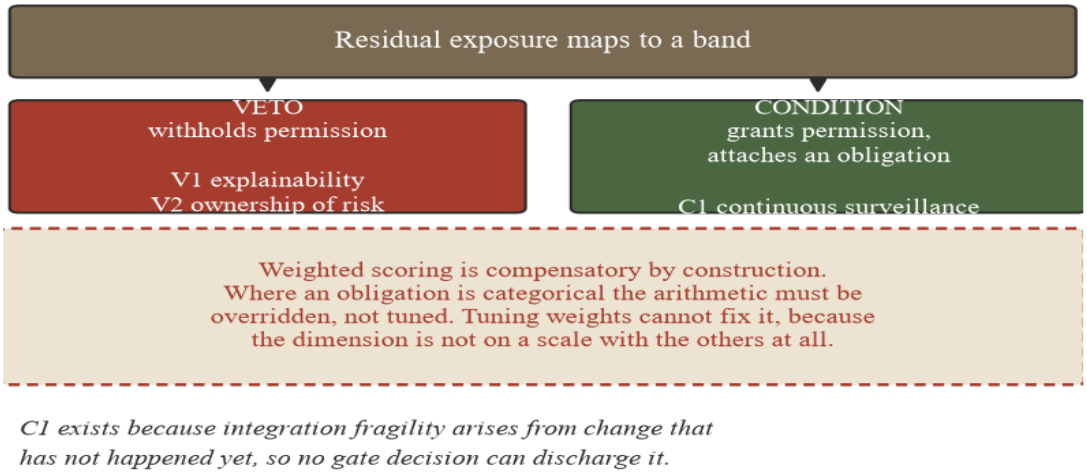

Figure 5. Why the arithmetic must be overridden rather than tuned. A categorical obligation is not on a scale with the other dimensions.

Veto rules withhold permission. V1 addresses explainability: where a covered decision can produce an adverse action, the law requires a specific and accurate statement of reasons, complexity is not a defence, and post-hoc approximations must themselves be validated. An explainability deficit in that setting is not a risk to be offset against other strengths. It is a condition precedent that has not been met, and no amount of benchmark testing, sample review,

or nondiscrimination testing cures an inability to tell a declined applicant why. V2 addresses ownership of risk: the valuation rule places responsibility on the institution using the model, not the one that built it, and an institution cannot discharge a duty to ensure confidence in estimates and test for nondiscrimination while lacking the documentation and testing rights required to do either.

Condition rules grant permission while attaching a lasting obligation. C1 differs from a veto in kind rather than degree. A veto withholds permission; a condition rule grants it while attaching an obligation that outlives the gate. It must exist because integration fragility cannot be discharged by anything decided at the moment of adoption: the risk arises from change that has not happened yet, so the only responsive control runs continuously. A framework treating every risk as resolvable at the gate would systematically understate this class of risk, and would do so invisibly.

7. Worked Illustration

The following is a synthetic construction. The lender does not exist, the tools do not exist, and every input score is invented to make the mechanics visible. The arithmetic is correct. Correct arithmetic on invented numbers is an illustration, not a measurement, and nothing in this section is a result.

Two tools of comparable sophistication are evaluated by the same institution and reach opposite outcomes, because of what the decisions are rather than what the technologies are.

Both land in the same band on residual exposure alone. A framework consisting only of weighted scoring would have treated them identically. The overrides separate them, and separate them in different directions for unrelated reasons. One is held back because a legal precondition is unmet, which is a question settled at the gate. The other is admitted but encumbered, because its dominant risk cannot be settled at the gate at all.

A third observation is retained deliberately even though it works against the framework. The second tool's strongest available control is the one least responsive to its largest risk. Complete human review earns the highest single credit in the catalogue and does almost nothing about silent field loss, because the reviewer and the tool share the same blind spot. A framework scoring controls purely by catalogue weight would have concealed that entirely. Separating D7 out and attaching C1 is what surfaces it. A framework that only ever demonstrates its own success is less credible than one that shows where its own arithmetic misleads.

8. RESULTS

There are none.

Table 4. Requested results against actual.

Requested	**Actual**
Tools evaluated	0
Approved	0
Adopted with mitigations	0
Rejected	0
Risks caught pre-production	0
Time added or saved	Not measured

Time added or saved Not measured

The framework has never been applied to a real tool. The examples in Section 7 are synthetic constructions with invented inputs, and their scores are arithmetic rather than observation. No results section is populated, because a results table in a permanently citable publication would be fabricated data.

9. Validation Studies That Would Establish the Framework

Four studies, in the order it would make sense to conduct them.

Inter-rater reliability. A panel of practitioners scores the same tool-and-decision pairs independently. The statistic that matters is agreement on band assignment rather than on individual dimension scores, because the band governs action: two assessors disagreeing on whether a dimension is 3 or 4 but agreeing the tool belongs in restricted pilot have not meaningfully disagreed. Low agreement would indicate the anchors are insufficiently specific, which is fixable. This is the cheapest study and the most informative about whether the instrument works at all.

Weight elicitation. Weights elicited from practitioners and regulators through pairwise comparison or discrete choice, compared against the stipulated values. Substantial divergence would indicate the current weights encode individual rather than shared professional judgement, which is a real possibility worth knowing.

Retrospective concordance. Applying the framework to adoption decisions whose outcomes are known, testing whether tools that caused problems would have scored into more restrictive bands. Vulnerable to hindsight bias; would require blinded scoring to mean anything.

Prospective comparison. Institutions using the framework against matched institutions using existing practice, measured on documentation quality, incidence of tool-attributed error, and time to adoption. The only design capable of establishing that the framework improves decisions, and by far the most expensive.

The first is realistically achievable by a single researcher with access to a professional network, and is the natural next piece of work.

10. What the Framework Assumes About the Institution

The framework is institution-agnostic in the sense that it does not require supervisory status. It nonetheless assumes several things about the adopting organisation, and where those assumptions fail the mechanism degrades in predictable ways.

It assumes someone can state how the tool will be used. Stage 1 requires a sentence describing the decision the tool will influence and in what sense. In an organisation where a tool is procured centrally and adopted opportunistically by different teams for different purposes, there is no single answer, and the framework's unit of analysis dissolves. The response is to assess per use rather than per tool, which multiplies the assessment burden in exactly the organisations least equipped to carry it.

It assumes controls credited will be implemented. Stage 3 credits only controls the institution commits to, with a named owner. The framework has no mechanism to verify that the named owner implemented anything, and a credited control that never materialises reduces residual exposure on paper while changing nothing. The named owner is a social device rather than a technical one, and it works to the extent that the record is read later.

It assumes the re-review trigger fires. Adoption records set triggers on vendor update, scope change, calendar interval, or surveillance breach. Whether anyone acts on a fired trigger is outside the framework, and a trigger that fires into an unmonitored queue is equivalent to no trigger.

It assumes the assessor is not the advocate. Nothing in the framework prevents the person proposing a tool from scoring it, and the scoring has enough judgement in it that an advocate would produce systematically lower inherent risk and higher control credit without any individual score being indefensible. Separating proposal from assessment is a governance requirement the framework depends on and does not state.

Table 5. Institutional assumptions and their failure modes.

Assumption	If it fails	Mitigation available
A single stated use per tool	The unit of analysis dissolves	Assess per use; cost multiplies

Credited controls get implemented	Residual exposure understates reality	Verify implementation before the record closes
Re-review triggers are acted on	Approved tools drift unmonitored	Route triggers to a monitored queue with an owner
Assessor independent of advocate	Systematic optimism in scoring	Separate the roles explicitly

The common feature is that each failure produces an assessment that looks complete. The framework's outputs are the same shape whether or not these assumptions hold, which means the record cannot distinguish a rigorous assessment from a performed one. That is a real limit, and it is shared with every governance instrument that produces a document.

11. How an Institution Would Actually Run This

The framework describes an assessment. Running it inside an institution raises questions the design does not address and a first adopter would face immediately.

Who assesses. Section 11 notes that nothing prevents the tool's advocate from scoring it. In practice the assessor needs enough domain knowledge to characterise the decision and enough independence to score it unfavourably, and those two requirements pull apart: the people who understand how a tool will be used are usually the people who want it.

When it runs. The framework is pre-adoption, which means it must be triggered by something. Procurement is the obvious trigger and it misses tools acquired without procurement, which per Section 3 is how the most dangerous adoptions happen. A trigger tied to procurement catches the tools that were already being handled carefully.

What it costs per assessment. Seven dimensions with written justification, a control catalogue with named owners, and a documented record is a matter of hours rather than minutes. That is proportionate for a consequential tool and disproportionate for a trivial one, and Section 11 acknowledges that the proportionality has not been demonstrated to calibrate correctly.

Where the record lives. The output is an artifact whose value is realised when someone asks a question years later. A record held in a project folder by the team that ran the assessment will not be findable then, and the framework's principal claimed benefit over ad-hoc evaluation depends entirely on retrieval.

None of these is a defect in the scoring model. They are the questions that determine whether the scoring model gets used, and they are organisational rather than methodological.

12. LIMITATIONS

These are stated fully rather than left for a reviewer to find.

The framework has no notion of aggregate exposure. Each assessment considers one tool in one decision. An institution adopting twenty tools each scoring into the second band has accumulated something the framework cannot express, since twenty individually acceptable exposures may not be collectively acceptable and nothing sums them.

Reassessment cadence is trigger-based only. Re-review fires on vendor update, scope change, calendar interval, or surveillance breach. A tool whose decision context drifts gradually, with no discrete event, is not caught by any of these, and gradual drift in how a tool is used is precisely the authority-through-use mechanism described in Section 3.

Tools that do not fit the dimensions. The framework evaluates a single tool influencing a single decision. It handles chains of tools acting on one another's outputs badly. Where one system's output becomes another's input without an intervening human, the risk is a property of the composition rather than of any component, and the seven dimensions do not capture that. Given where adoption is heading, this is the most pressing extension.

Regulations change. The framework is built against United States federal law as it stands. State-level obligations are not modelled, and anything outside the United States would need substantial adaptation rather than translation. The ground is actively moving, and a framework anchored to a regulatory moment inherits that moment's shelf life.

Rigour against speed, unresolved. A framework imposing uniform burden gets circumvented for low-stakes tools, and once circumvented for those it loses credibility for the high-stakes ones that were the only ones that mattered. Proportionality is the intended answer, and it has not been demonstrated that the proportionality calibrates correctly. Whether the framework is light enough on low-stakes tools to avoid being routed around is an empirical question nobody has answered.

When an approved tool later fails. Partially addressed by re-review triggers and C1, and only partially. C1 delegates a risk to a monitoring regime whose adequacy the framework never assesses. An institution can satisfy C1 with entirely nominal surveillance: an alert nobody reads, a threshold set so wide it never fires. The framework can require monitoring. It cannot require the monitoring to be good.

Stipulated values. Weights, credits, and thresholds are stipulated rather than estimated. Different values produce different scores and, near a boundary, different outcomes.

Ordinal scores treated as interval. The framework performs arithmetic on ordinal scores as though they were interval-scaled. This is a well-known and well-criticised property of weighted scoring models generally. The defence is narrow and should be stated as narrowly as it deserves: the alternative currently in use is unstructured judgement, which has the same defect without the transparency that makes it contestable. That is a defence of the framework relative to current practice. It is not a defence of the arithmetic.

Band discontinuities. A residual exposure of 24.8 and one of 25.2 are not meaningfully different, yet receive different outcomes. The near-boundary rule addresses this and an institution adopting the framework should write the convention down explicitly.

Control credits are stipulated with less basis than the weights. The dimension weights at least reflect a stated reasoning about where consequence concentrates. The individual control credits reflect a judgement about how strongly the evidence supports each control addressing its named failure mode, and that judgement is less articulable. The relative ordering is more defensible than the values, and an institution adopting the framework should expect to revise them.

The framework assumes a single accountable assessor exists. Stage 4 requires an accountable signatory. In organisations where tool adoption is genuinely distributed, identifying who that is may be harder than any of the scoring, and the framework offers no guidance on it.

No inter-rater reliability data. Two assessors scoring the same tool may reach different conclusions and it is not known by how much.

13. FUTURE WORK

The inter-rater study, before anything else. If practitioners cannot apply the anchors consistently, none of the framework's other properties matter.

A usable decision record template. The record structure is described in prose but not built as a usable form. Producing one would be straightforward and genuinely useful, since it is the artifact an institution would actually hold in its hands.

Extension to tool chains. As Section 10 notes, composition risk is the most pressing gap given the direction of adoption.

Calibration against the forthcoming regulatory position. The agencies have signalled a request for information on banks' use of AI including generative and agentic systems, which could change the analysis materially.

A note on what the first study would settle. Inter-rater reliability is listed first not because it is easiest but because it is prior to everything else. If practitioners cannot apply the anchors consistently, then weight elicitation measures disagreement about a scale nobody applies the same way, retrospective concordance measures a scorer rather than a framework, and prospective comparison compares two organisations using different instruments that share a name.

It is also the study whose negative result is most actionable. Low agreement on band assignment points at the anchors in the appendix, which are the most straightforwardly revisable part of the design. A framework that fails inter-rater reliability and is fixed by writing better anchors has been improved; a framework that skips the study retains an unknown defect in the part everything else depends on.

14. CONCLUSION

Adoption of an AI tool into a regulated financial decision is frequently not made as a risk decision. It is made as a procurement decision, or it happens through pilot drift where the tool acquires authority without anyone granting it. Governance engages afterwards, which is structurally too late for obligations that attach at first use.

The framework proposed here engages before that point, takes the decision rather than the tool as its unit of analysis, and produces a documented artifact where previously there was only reasoning that lived in conversation. Its principal design contribution is not the scoring but the overrides: weighted additive scoring is compensatory by construction, some legal obligations are categorical, and where those meet, the arithmetic has to be overridden rather than tuned. The condition rule exists because one class of risk, silent degradation after an upstream change, cannot be discharged by anything decided at the moment of adoption.

The modest claim is the one worth ending on. Structure does not make a judgement correct. It makes it inspectable, and inspectability is what an ad-hoc evaluation cannot provide however careful it was.

Nothing here has been tested. No tool has been assessed, no institution has used the framework, and the weights and thresholds are stipulated values reflecting one person's reading of where consequence concentrates. Whether applying this produces better adoption decisions than the practice it would replace is a hypothesis, and Section 9 states the four studies that would begin to answer it. The first of them is cheap, and until it is run the framework's most basic property, that two assessors would reach the same band, is unknown.

REFERENCES

1. Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, Supervisory Guidance on Model Risk Management, SR 11-7 / OCC 2011-12, April 2011.
2. Consumer Financial Protection Bureau, Regulation B, Equal Credit Opportunity Act, 12 CFR 1002.9, Notifications.
3. Consumer Financial Protection Bureau, Consumer Financial Protection Circular 2022-03: Adverse Action Notification Requirements in Connection with Credit Decisions Based on Complex Algorithms, May 2022.
4. Federal financial regulatory agencies, Quality Control Standards for Automated Valuation Models, interagency final rule, 2024.
5. 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.