TOLERANCE FOR OPACITY

A Threshold Framework for AI-Driven Banking

JM García-Maceiras

President of the Spanish BPO Banking Association

Title: *Tolerance for Opacity: A Threshold Framework for AI-Driven Banking*[1].

Abstract—The increasing deployment of artificial intelligence in banking raises a structural tension between algorithmic opacity and the institutional obligation to justify decisions under supervisory, judicial, and systemic scrutiny. While technical complexity may enhance predictive performance, it can simultaneously weaken the capacity of financial institutions to reconstruct and defend decision-making pathways when formally challenged.

This article introduces the concept of *Tolerance for Opacity* (TfO) as a threshold framework for assessing the level of opacity compatible with stable institutional reconstructibility. *Reconstructibility* is understood as the institutional capacity to ex post rearticulate the causal, normative, and evidentiary chain underlying an AI-assisted decision in a form that remains defensible under conditions of formal scrutiny. The paper argues that opacity does not become problematic merely because it limits interpretability; it becomes prudentially relevant when it approaches a level at which reconstructibility may degrade abruptly.

The analysis develops the threshold problem conceptually rather than quantitatively. It proposes a structured governance map to visualize the interaction between aggregate opacity and institutional capacity, identifying zones of stability and instability within AI-driven banking environments. The objective is not to eliminate opacity, but to bound its institutional consequences.

Situated in the context of the forthcoming implementation of the EU Artificial Intelligence Act and its interaction with existing financial supervisory obligations, the framework seeks to clarify how prudential architecture must evolve to preserve accountability under increasing technological complexity.

---

[1] This work continues the research project initiated in *García-Maceiras, JM (2026). "The Banking Risk of AI Explanation". ADR Notebooks No. 1. Zyphrum Alchemists.* ISBN 9789403845760; and continued with *García-Maceiras, JM (2026). "The Five Beacons Model: A Prudential Architecture for AI Explainability and Legal Liability in Banking"* DOI 10.5281/zenodo.18647317. It concludes with the forthcoming: "*Systemic Opacity Risk*".

Keywords: *Artificial Intelligence in Banking*; *Explainability; Algorithmic Opacity*; *Reconstructibility*; *Tolerance for Opacity*; *Model Risk Governance*; *Supervisory Accountability*; *AI Governance*; *Prudential Regulation*; *Algorithmic Decision-Making.*

---

1. Introduction: From Explainability to Institutional Reconstructibility

The rapid integration of artificial intelligence systems into core banking functions—credit assessment, capital allocation, liquidity management, fraud detection, customer interaction—has intensified the regulatory and academic debate on AI explainability. Much of this debate has focused on the interpretability of models, the transparency of algorithms, and the intelligibility of outputs. These discussions have been essential in clarifying the technical challenges of Black-box architectures and the limits of human interpretability in high-dimensional systems[2].

However, in prudentially regulated environments, the central question is not exhausted by whether a model can be explained. The more structural issue is whether an institution can remain accountable for the decisions produced by its AI systems when subjected to formal scrutiny. Supervisory review, judicial proceedings, internal audit, and fiduciary obligations all presuppose that automated decisions can be retraced in a manner sufficient to assess compliance, proportionality, and procedural integrity. In this context, the relevant problem is institutional rather than purely technical.

Existing approaches to explainable AI are predominantly model-centric. They seek to render model behavior interpretable through feature attribution techniques, surrogate models, saliency maps, or *post hoc* explanations. Governance frameworks, by contrast, are often compliance-centric, focusing on documentation, validation processes, and internal controls. What remains under-theorized is the institutional condition that connects these two domains: the capacity of a regulated entity to retrace and justify the pathway by which an automated system transformed inputs into a specific decision under conditions of formal scrutiny.

This paper proposes *reconstructibility* as that missing institutional variable. Reconstructibility does not demand full algorithmic transparency, nor does it presuppose complete human intelligibility of complex models. Rather, it denotes the structured institutional capacity to retrace a decision pathway capable of causal articulation when exposed to supervisory, judicial, or internal examination, at a level sufficient to discharge regulatory and fiduciary responsibilities. It is therefore a condition of accountability, not a claim about the eliminability of technical complexity.

The argument builds upon the architecture previously advanced in The Five Beacons Model (5B), which identified structural orientation points for the governance of AI in banking. Those *Beacons* were conceived as institutional reference markers capable of guiding responsible AI deployment without imposing technological prohibitions. The

[2] For foundational discussions on interpretability and its limits, see Doshi-Velez & Kim (2017); Lipton (2018); Rudin (2019).

present framework develops the prudential condition under which such orientation remains effective. Where automated systems cannot be retraced under scrutiny, the guiding function of governance principles is weakened; where decisions can be institutionally re-illuminated, accountability retains structural stability even in environments of high computational complexity.

The focus of this paper is deliberately confined to banking. This sector offers a uniquely dense regulatory environment in which supervisory review, capital adequacy assessment, and litigation exposure converge. If a concept of institutional reconstructibility can be operationalized within this setting, it may serve as a robust prudential tool without requiring the introduction of new regulatory categories or abstract (or paramount) transparency mandates. The ambition is therefore not to prohibit algorithmic opacity, but to calibrate its tolerable limits within an accountable institutional architecture.

From this foundation emerges the notion of *Tolerance for Opacity* (TfO). Rather than treating AI opaqueness as a binary defect, TfO conceptualizes it as a graduated condition. Regulated entities inevitably operate with models of varying complexity and partial non-traceability. The relevant prudential question is not whether technical opacity exists, but how much of it can be institutionally managed without impairing reconstructibility. TfO thus denotes the maximum level of institutionally managed opacity that a banking institution may assume without compromising its ability to justify its AI-driven decisions under formal scrutiny.

The remainder of the text develops this argument in four steps: first, it reframes algorithmic opacity in prudential terms as *the degradation of reconstructibility*; second, it specifies *reconstructibility* conceptually and normatively; third, it proposes a minimalist taxonomy of opacity—Model Opacity, Process Opacity, and Accountability Opacity—designed for supervisory applicability; finally, it introduces the conceptual foundation of *Tolerance for Opacity* as a calibrable prudential threshold, whose metric elaboration follows in subsequent sections.

The present analysis intentionally privileges structural coherence over exhaustive doctrinal mapping. It seeks to articulate a conceptual architecture capable of guiding institutional design, rather than to catalogue the full spectrum of AI governance literature; it does not seek to exhaust the conceptual space of opacity governance. The goal is not to measure but to define conditions of possibility.

This perspective builds upon the layered AI governance architecture previously articulated in *The Five Beacons Model* (5B), where differentiated accountability functions were structured as complementary safeguards against institutional opacity. TfO develops one specific dimension of that architecture: the calibration of opacity within prudentially sustainable limits.

2. Algorithmic Opacity: General Notion and Prudential Relevance

2.1 Opacity in the Contemporary Debate

Algorithmic opacity has become a central concern in discussions on artificial intelligence, particularly in contexts involving automated decision-making. In technical literature, opacity is commonly associated with Black-box architectures, high-dimensional parameter spaces, and the difficulty of interpreting non-linear model behavior. In legal and policy discourse, AI opacity is often framed as a threat to transparency, fairness, and accountability.

These approaches, while valuable, tend to operate at two distinct but incomplete levels. On the one hand, technical research focuses on interpretability methods aimed at rendering model behavior more intelligible—through feature attribution, surrogate modelling, or local explanation techniques. On the other hand, governance frameworks emphasize documentation requirements, validation procedures, and internal controls as mechanisms for oversight[3].

Both perspectives are necessary. Yet neither fully captures the institutional dimension of opacity in regulated sectors such as banking. Technical opacity does not automatically translate into regulatory failure; nor does compliance documentation necessarily ensure meaningful accountability. A model may remain complex while being institutionally manageable, and conversely, a formally compliant system may still be irretrievable under scrutiny.

In prudential environments, AI opacity becomes normatively relevant not merely when a model is difficult to interpret, but when the institution deploying it cannot adequately account for its decisions under conditions of formal review.

2.2 From Technical Opacity to Institutional Opacity

Opacity may arise from multiple sources: the intrinsic complexity of AI model architecture, the scale and heterogeneity of data inputs, the inscrutability of training processes, the integration of external vendors and distributed infrastructures. However, from a supervisory standpoint, the relevant question is not the existence of complexity as such, but its institutional consequences.

In the banking context, automated decisions affect capital allocation, credit approval, risk weighting, liquidity planning, and customer treatment. These decisions are subject to layered scrutiny: internal audit, supervisory examination, regulatory reporting, judicial review, and, in some cases, public accountability. Under such conditions, algorithmic opacity is problematic when it impairs the institution's capacity to retrace how a given output was produced from specified inputs within an identifiable decision structure.

[3] On algorithmic opacity and the limits of transparency-based governance, see Burrell (2016); Ananny & Crawford (2018).

This shift—from model-centric opacity to institution-centric opacity—is critical. The prudential system does not demand that every parameter be intuitively comprehensible to human operators; it requires that the institution retain the structured capacity to justify its decisions in legally and regulatorily intelligible terms.

Opacity, in this sense, is not defined by the absence of transparency *per se*: it is defined by the *degradation of reconstructibility*.

2.3 Opacity as Degradation of Reconstructibility

Reconstructibility, as developed in this paper, refers to the institutional capacity to retrace an algorithmic decision pathway capable of causal articulation under formal scrutiny. It does not presuppose full interpretability of every computational layer, nor does it mandate disclosure of proprietary code. Rather, it concerns the stability of institutional accountability in the presence of AI complexity.

Opacity becomes prudentially significant when it reduces this capacity below a threshold compatible with supervisory, judicial, or fiduciary obligations. A system may remain technically opaque yet institutionally reconstructible if adequate documentation, logging, version control, and governance structures allow its decisions to be retraced in a structured and defensible manner. Conversely, even moderately complex systems may become institutionally opaque where implementation practices, outsourcing arrangements, or deficient controls render reconstruction impracticable.

This understanding reframes algorithmic opacity as a gradient condition rather than a binary defect. Institutions operate within environments of unavoidable complexity. The relevant prudential inquiry is therefore not whether opacity exists, but how much AI opacity can be sustained without undermining the reconstructive capacity that anchors responsibility.

In this perspective, opacity is neither an abstract ethical concern nor a purely technical characteristic. It is a structural variable affecting the resilience of institutional accountability. Where reconstructibility is preserved, complexity remains governable; where reconstructibility erodes, institutional responsibility becomes unstable—even if predictive performance remains high.

Under this light, *Tolerance for Capacity* (TfO) is not merely an internal management ratio; it functions as the threshold variable that defines the boundary of supervisory acceptability. Beyond this point, any accumulation of opacity—whether technical or procedural—effectively dissolves the institutional capacity for accountability, rendering the AI system's outputs indefensible.

2.4 Prudential Specificity

The reframing of algorithmic opacity in terms of reconstructibility is particularly pertinent in banking. Prudential regulation is fundamentally oriented toward stability, traceability of risk, and the capacity to justify capital and liquidity positions under

scrutiny. Supervisory processes such as on-site inspections, model validation reviews, and stress testing implicitly rely on the assumption that institutional decisions can be retraced and assessed.

In this environment, opacity is not problematic because it offends an abstract transparency ideal; it is serious when it compromises the institution's ability to demonstrate compliance, reconstruct causal chains in contested decisions, or justify risk-weighted asset calculations under review. The prudential relevance of AI opacity thus derives from its impact on accountability structures rather than from its technical form alone.

This institutional orientation allows for a more precise calibration of concerns. Not all forms of algorithmic opacity are equally significant. What matters is whether they degrade reconstructibility to a degree that affects regulatory oversight and fiduciary responsibility.

3. Reconstructibility: Conceptual Specification and Normative Foundation

3.1 Canonical Definition

For the purposes of this framework, *Reconstructibility is the institution's capacity to ex post retrieve, articulate, and substantiate the causal, normative, and evidentiary chain underlying an AI-assisted decision in a form that remains institutionally defensible under supervisory, judicial, or systemic scrutiny, including in conditions of formal challenge.*

Several elements of this definition warrant clarification. Reconstructibility is institutional rather than individual. It does not refer to the cognitive capacity of a single engineer or compliance officer to interpret model internals, but to the organizational ability of the institution to retrace, document, and justify—and, where necessary, rearticulate—a specific decision when formally required to do so.

As previously argued in the *The Five Beacons Model* governance framework, reconstructive capacity is not a spontaneous attribute of complex systems but the outcome of institutional design choices that distribute AI documentation, oversight, and accountability functions across organizational layers.

Retracing does not imply full algorithmic transparency. Complex AI systems may involve layers of computation that resist intuitive interpretability. Reconstructibility requires not exhaustive intelligibility, but structured retrievability of the decision pathway to a degree that permits meaningful scrutiny.

The framework does not assume that full epistemic transparency of complex AI systems is always attainable. It proceeds from a more modest premise: that institutions remain normatively obligated to preserve sufficient reconstructive capacity to justify decisions under conditions of stress. Tolerance for Opacity (TfO) does not eliminate opacity; it bounds its prudential consequences.

The standard is relational and context-sensitive. The "level of granularity sufficient" is not abstractly defined but linked to the nature of the obligations at stake—supervisory review, litigation exposure, capital justification, fiduciary accountability. Reconstructibility is therefore a graded institutional capacity, not a binary attribute. In litigation-intensive contexts, this relational standard ultimately determines the decision's forensibility[4].

3.2 Normative Foundation

Reconstructibility is not advanced as a general transparency ideal, nor as a demand for full algorithmic interpretability. Its normative foundation is narrower and institutional.

In regulated environments such as banking, the legitimacy of automated decision-making depends on the capacity of the institution to account for its decisions under conditions of formal scrutiny. Supervisory examination, judicial assessment, and fiduciary responsibility presuppose that decisions can be retraced in a manner sufficient to evaluate compliance, proportionality, and procedural integrity. Where such retracing becomes structurally impossible, institutional accountability is impaired, regardless of the predictive performance of the underlying model.

In this sense, reconstructibility represents the institutional capacity to re-illuminate algorithmic decisions when exposed to scrutiny, preserving the decision's forensibility under adversarial examination. It is not a claim about technological simplification, but about the preservation of responsibility within increasingly complex decision environments. The prudential concern arises not from AI opacity as such, but from the lack of transparency that destabilizes the structures through which responsibility is exercised and reviewed.

3.3 Scope and Delimitation

The concept of *reconstructibility* must be carefully delimited to avoid conceptual inflation. Reconstructibility does not claim novelty as a technical term; it has been employed in other domains (particularly in control theory and systems analysis) to describe the capacity to recover internal system states from observable outputs. The present framework adapts the term to the institutional governance of AI-driven decision-making, where the relevant object of recovery is not merely a mathematical state but an accountable decision pathway.

Reconstructibility is not equivalent to *explainability*. Explainability techniques aim to render model behavior intelligible to human observers, often through approximations or local explanations. *Reconstructibility*, by contrast, concerns the institutional ability to retrace the specific pathway that led to a particular decision in a manner defensible under formal scrutiny.

[4] The term is used descriptively to refer to the evidentiary dimension of reconstructibility and does not introduce a separate normative category.

Nor is *reconstructibility* identical to model *accuracy* or *robustness*. A model may perform optimally according to statistical metrics while remaining difficult to retrace institutionally. Conversely, a moderately complex model may be highly reconstructible if supported by robust logging, documentation, and governance structures.

Reconstructibility should not be conflated with *auditability* or *traceability*. *Auditability* typically refers to the availability of logs or technical traces. *Reconstructibility*, by contrast, denotes an institutional capacity: the structured ability to reassemble, *ex post*, the causal, normative, and evidentiary chain of a decision in a manner capable of sustaining adversarial scrutiny. It is therefore not merely procedural but architectural. *Reconstructibility* is not a synonym of *accountability*; it is the institutional precondition for its credible exercise[5].

Importantly, *reconstructibility* does not require public disclosure of proprietary code or unrestricted transparency. It is compatible with intellectual property protection and commercial confidentiality. The relevant audience is not the general public, but those actors legitimately entitled to scrutinize the decision: supervisors, courts, and authorized internal bodies.

3.4 Reconstructibility as a Gradient Condition

Reconstructibility is inherently graduated. Institutions may possess higher or lower degrees of reconstructive capacity depending on model complexity, implementation practices, governance arrangements, and documentation quality.

This gradation is essential for prudential calibration. If reconstructibility were treated as an all-or-nothing requirement, complex AI systems would be structurally excluded from regulated environments. Conversely, if reconstructibility were reduced to minimal formal documentation, it would lose its normative force.

The relevant question is therefore not whether reconstructibility exists in absolute terms, but whether it is preserved above a threshold compatible with supervisory, judicial, and fiduciary obligations.

The viability of any reconstructibility-based framework depends on the prior existence of a structured AI governance architecture. Reconstructibility is not an emergent property of complex systems; it is the product of deliberate institutional design. The present paper therefore assumes, without restating in full, a layered governance model that distributes accountability, documentation, and oversight functions across organizational strata.

[5] On accountability as structured answerability within institutional systems, see Bovens (2007).

4. A Minimalist Taxonomy of Prudential Opacity

4.1 Rationale for a Minimalist Approach

If *opacity* is understood as *degradation of reconstructibility,* it becomes necessary to identify the principal loci within which such degradation may occur. A comprehensive philosophical taxonomy of opacity would distinguish epistemic, ontological, cognitive, and communicative dimensions. Such granularity, however, risks diluting prudential applicability. These domains are analytically distinct, yet their governance implications depend on how institutional accountability layers are structured and coordinated.

The present framework adopts a deliberately minimalist approach. Its objective is not to catalogue all forms of algorithmic opacity, but to isolate those structural domains in which reconstructibility may be impaired in regulated banking environments. The taxonomy therefore prioritizes supervisory operability, conceptual clarity, and institutional relevance over theoretical exhaustiveness.

Three principal domains are identified: Model Opacity, Process Opacity, and Accountability Opacity[6]. These categories do not assume full independence; rather, they mark distinct institutional loci at which *reconstructibility* may degrade.

4.2 Model Opacity (MO)

Model Opacity refers to the degree to which the internal computational architecture of an AI system resists structured retracing under scrutiny.

This facet concerns characteristics inherent to the model itself, including but not limited to high-dimensional non-linear structures; deep neural networks with distributed parameter representations; ensemble systems with layered decision aggregation; architectures whose internal states cannot be directly mapped onto interpretable variables[7].

Model Opacity does not equate to model risk in the conventional prudential sense. Model risk typically concerns inaccuracy, instability, bias, or performance degradation. A model may be statistically robust and empirically validated while remaining structurally obscure in terms of reconstructibilty. Conversely, a relatively simple model may exhibit performance deficiencies without being unfathomable in reconstructive terms.

Model Opacity becomes prudentially significant when architectural complexity materially impairs the institution's ability to articulate the decision pathway under

[6] *Communicative capacity* may be understood as the institution's ability to translate reconstructible decision pathways into intelligible explanations addressed to supervisors, courts, or affected clients. It does not constitute an independent opacity category but operates transversally across Model, Process, and Accountability domains. Its impairment may aggravate reconstructibility fragility without altering underlying architectural opacity.

[7] For discussions on transparency and automated decision-making, see Zarsky (2013); Wachter et al. (2017).

formal review. The issue is not the existence of complexity, but whether such complexity (and intensity[8]) undermines reconstructibility beyond acceptable thresholds.

Importantly, Model Opacity is not necessarily eliminable. Certain AI architectures may deliver demonstrable performance benefits while retaining intrinsic structural opacity. The prudential question is therefore not whether Model Opacity can be reduced to zero, but how it interacts with other institutional safeguards.

4.3 Process Opacity (PO)

Process Opacity concerns the operational environment within which the model is developed, implemented, updated, and monitored. Even a technically interpretable model may become institutionally enigmatic if implementation practices impair retrievability.

This dimension includes multiple factors—inadequate logging of inputs and outputs, deficient version control, insufficient documentation of model updates or retraining cycles, integration of data sources lacking auditability; weak change-management procedures.

Process Opacity is typically more directly manageable than Model Opacity. It reflects the quality of internal controls, documentation practices, and operational governance structures. Failures in this domain may render reconstructibility impracticable even where model architecture would otherwise permit retracing.

From a prudential standpoint, Process Opacity is particularly salient because it falls squarely within the institution's sphere of control. Where reconstructibility fails due to operational deficiencies, the impairment is not attributable to technological inevitability but to governance weakness.

4.4 Accountability Opacity (AO)

Accountability Opacity arises from the institutional allocation (or fragmentation) of responsibility for AI-driven decisions.

In increasingly complex banking ecosystems, AI systems may be developed by external vendors, hosted on third-party infrastructure, integrated across multiple organizational units, subject to overlapping governance frameworks.

Where responsibility for model design, deployment, monitoring, and validation is dispersed without clear attribution, reconstructibility may be structurally compromised. An institution may possess technical access to outputs and documentation, yet lack a coherent chain of responsibility capable of sustaining formal scrutiny.

---

[8] *Opacity intensity* reflects the degree of structural depth and adaptive complexity within a model architecture. It differs from mere presence of opacity; intensity captures the magnitude of computational layering and state evolution that may complicate reconstructibilty.

Accountability Opacity therefore concerns the institutional architecture within which AI operates. It addresses questions such as: *Who is ultimately responsible for the decision? Who can authoritatively retrace and justify it? Does the institution retain sufficient access to underlying processes in outsourced arrangements? Are escalation and oversight mechanisms clearly defined?*

This dimension is distinct from purely contractual or corporate governance issues. It relates specifically to the stability of the responsibility structure that underpins reconstructibility. Where accountability chains are fragmented or ambiguous, even technically retrievable systems may become institutionally opaque.

4.5 Interactions and Non-Independence

The three dimensions (MO, PO, AO) are analytically distinguishable but not fully independent. High Model Opacity may increase reliance on robust operational documentation to preserve reconstructibility. Weak process controls may amplify the impact of architectural complexity. Fragmented accountability structures may undermine the effective use of otherwise adequate logs and documentation.

The taxonomy does not assume additive separability in a strict mathematical sense. Rather, it identifies structurally distinct points of potential degradation. Any assessment of opacity must therefore consider interaction effects and contextual dependencies.

5. Limits and Structural Vulnerabilities of the Proposed Taxonomy

This taxonomy is intentionally limited in scope. It does not attempt to classify epistemic opacity in philosophical terms, nor does it address broader societal or ethical concerns beyond prudential accountability. Its function is to provide a supervisory-friendly framework for identifying where reconstructibility may be impaired within regulated banking institutions.

By isolating three principal domains, the taxonomy seeks to avoid conceptual inflation while preserving analytical precision. Its adequacy, however, depends on careful calibration and awareness of its structural limitations. This section identifies structural vulnerabilities that may affect the robustness of the proposed taxonomy.

5.1 Boundary Instability

In practice, the three dimensions are not cleanly separable. For example, inadequate documentation of model retraining may be characterized as Process Opacity (PO). Yet if retraining materially alters internal representations in a manner that renders prior behavior irrecoverable, Model Opacity (MO) is simultaneously implicated. Similarly, outsourcing core model components to a third-party provider may appear as Accountability Opacity (AO), while limiting access to internal parameters and thereby increasing Model Opacity (MO).

The taxonomy assumes analytical distinction, not ontological separation. Under real supervisory examination, these domains may collapse into one another. If the framework is applied rigidly, it risks misclassification or artificial compartmentalization of what are, in effect, structurally intertwined phenomena.

5.2 Reductionism Risk

By limiting opacity to three prudential domains, the taxonomy necessarily excludes broader epistemic, ethical, and socio-technical dimensions. Opacity may arise not only from architectural complexity, operational deficiencies, or fragmented accountability, but from data asymmetries, feedback loops, adaptive drift in dynamic systems, or strategic opacity introduced by commercial actors.

The minimalist structure is defensible only if it is acknowledged as purpose-bound. It is not a general theory of algorithmic opacity. Its function is confined to preserving reconstructibility in regulated banking environments. Should it be applied outside that scope without adaptation, conceptual distortion may occur.

5.3 Measurement Ambiguity

The taxonomy identifies domains of opacity but does not provide a metric. Without a structured method of calibration, classification risks remaining descriptive rather than operational.

The transition from qualitative identification (e.g., "high MO") to prudential threshold determination requires additional analytical tools. Absent such tools, supervisory application may revert to discretionary judgment without consistent benchmarks. The taxonomy therefore depends on subsequent formalization—specifically, on the development of a calibrated threshold framework.

5.4 Dynamic System Evolution

AI systems are not static artefacts. Continuous learning, periodic retraining, and evolving data environments may alter the opacity profile of a system over time. A model initially classified as moderately opaque may become substantially more difficult to retrace following iterative updates. Process controls that were adequate at deployment may degrade in practice. Accountability chains may fragment as organizational structures evolve.

The taxonomy, as presently formulated, is structurally static. Without embedding it within ongoing monitoring and reassessment mechanisms, reconstructibility may erode gradually without triggering explicit institutional response.

5.5 Strategic Gaming and Formal Compliance

Any prudential framework creates incentives for strategic alignment. Institutions may respond to supervisory expectations by optimizing formal compliance indicators rather than substantive reconstructibility.

For instance, extensive documentation may be produced without ensuring meaningful artifiligence retrievability. Accountability structures may be formally clarified while practical access to relevant technical expertise remains limited. Model simplifications may be introduced solely to meet reconstructibility thresholds at the expense of performance or risk sensitivity. The taxonomy cannot, by itself, prevent such gaming. Its effectiveness depends on supervisory judgment and institutional integrity.

5.6 Structural Dependence on Reconstructibility

The taxonomy is conceptually dependent on reconstructibility as its organizing principle. If reconstructibility were challenged as an inadequate or incomplete normative anchor (on the grounds that accountability may require more than reconstructibility) the taxonomy would require reconsideration.

Critics may argue that causal retracing does not guarantee fairness; reconstructibility does not ensure non-discrimination; accountability may demand participatory or distributive dimensions beyond institutional reconstructibilty. These objections do not invalidate the framework but delimit its scope. The taxonomy is constructed to preserve institutional accountability within prudential structures. It does not purport to resolve all normative concerns surrounding AI deployment—far from it.

5.7 Functional Justification Despite Vulnerabilities

Attempts to elaborate a maximalist taxonomy have yielded us no discernible gains in terms of reconstructibility, institutional defensibility, or systemic stability. Excessive classificatory granularity, beyond what is required for evidentiary articulation, risks obscuring rather than enhancing the capacity of courts or supervisors to assess responsibility in AI-assisted decisions.

The existence of vulnerabilities does not undermine the utility of the minimalist taxonomy. Rather, it clarifies its epistemic status: it is a supervisory instrument, not an ontological cartography of Opacity.

Its adequacy should therefore be assessed pragmatically. If it enables supervisors and institutions to identify and calibrate threats to reconstructibility with sufficient clarity and consistency, it fulfils its purpose. Its limitations are structural consequences of its minimalist design.

6. Tolerance for Opacity (TfO)

We must move now from structural classification to prudential calibration. That transition requires defining the institutional threshold at which opacity ceases to be tolerable. At this point the concept of *Tolerance for Opacity* emerges as a measurable prudential variable.

6.1 From Opacity Domains to Threshold Logic

We have identified the principal domains in which opacity may degrade reconstructibility and have acknowledged the structural limits of the proposed taxonomy. However, classification alone is insufficient for prudential governance.

Supervisory practice does not operate at the level of descriptive mapping; it operates at the level of thresholds. The operative question is therefore not merely *Where does opacity arise?* but rather *At what point does opacity exceed the level compatible with institutional accountability?*

Tolerance for Opacity (TfO) addresses this threshold question. It introduces a prudential variable that connects structural opacity to supervisory sustainability. Opacity, in this framework, is neither categorically prohibited nor normatively neutral: it is institutionally tolerable only to the extent that reconstructibility remains intact above a defined accountability threshold.

6.2 Canonical Definition of TfO

In prior work, *Tolerance for Opacity* (TfO) was framed in descriptive terms as a governance-boundary concept within a prudential banking architecture. The present paper advances that initial formulation by elevating TfO to a formal analytical variable, enabling quantitative calibration and cross-institutional comparability. This shift does not modify the original intuition; rather, it refines its structural properties.

Tolerance for Opacity (TfO) may be defined as *the maximum aggregate level of institutional opacity that a regulated entity may assume without materially impairing its capacity to preserve reconstructibility under supervisory, judicial, or fiduciary scrutiny.*

Several components of this definition require emphasis. TfO is aggregate: it does not concern isolated instances of MO, PO or AO; it addresses the combined effect of these dimensions on reconstructibility. TfO is institution-specific: different institutions may operate with varying degrees of technical sophistication, governance robustness, and risk appetite; prudential calibration must therefore account for contextual institutional capacity.

TfO is dynamic: it may vary over time as models evolve, operational controls strengthen or weaken, and regulatory expectations shift. TfO is threshold-based rather than optimization-based: it does not aim to maximize opacity within permissible bounds, nor

to eliminate opacity entirely. It defines a boundary condition—outside this perimeter, reconstructibility becomes structurally unreliable.

6.3 The Relationship Between *Opacity* and *Reconstructibility*

Conceptually, TfO presupposes an inverse relationship between *opacity* and *reconstructibility*. As Model, Process, or Accountability Opacity increase, the institutional effort required to preserve reconstructibility intensifies. Up to a certain point, governance mechanisms (logging, validation, oversight structures) can compensate for architectural complexity. Beyond that point, compensatory mechanisms become insufficient.

The prudential threshold is reached when marginal increases in AI opacity produce disproportionate degradation in reconstructibility. At this juncture, decision pathways cannot be reliably retraced, capital justifications shift unstable, litigation defense becomes structurally impaired, and supervisory explanations lose substantive coherence. TfO therefore represents the maximum opacity level compatible with stable re-illumination of decisions under scrutiny.

It is important to underline that T*olerance for Opacity* does not logically derive from reconstructibility in a circular manner. Reconstructibility constitutes an institutional capacity (organizational, technical, evidentiary) that can be strengthened, weakened, or miscalibrated independently of opacity levels.

TfO expresses the maximum opacity-to-capacity ratio that an institution may assume without jeopardizing that capacity under stress conditions; TfO operates as a prudential boundary condition imposed upon opacity in light of that pre-existing capacity. The relationship is therefore asymmetrical: opacity tests reconstructibility; TfO constrains opacity to preserve reconstructibility.

6.4 Institutional Calibration Logic

To operationalize TfO, opacity must be assessed along the three identified domains (MO, PO, AO). These domains may interact non-linearly. High Model Opacity may be tolerable where PO and AO remain low. Conversely, moderate opacity across all three domains may collectively push the system beyond its reconstructibility threshold.

The calibration problem is therefore multidimensional. *Tolerance for Opacity* cannot be inferred from any single variable. It emerges from the interaction of the opacity profile with the institution's governance capacity. This implies that TfO is neither a purely technical metric nor a purely legal one. It is a prudential construct situated at the intersection of model architecture, operational governance, and responsibility allocation.

Institutional calibration of TfO is not a unilateral managerial prerogative. It remains subject to supervisory review, judicial assessment, and, where applicable, regulatory specification. TfO is therefore not a mechanism for legitimizing opacity, but a structured articulation of its permissible limits within externally reviewable constraints.

6.5 Prudential Function of TfO

The primary function of TfO is stabilizing. It does not seek to define optimal model design. It does not replace existing model risk management frameworks. It does not introduce a new regulatory category. Instead, it provides a structured lens through which supervisors and institutions may assess whether AI deployment remains compatible with the preservation of institutional accountability.

In this sense, TfO performs three prudential functions: (a) prevention, for it discourages excessive opacity accumulation before reconstructibility collapses; (b) diagnosis, it enables identification of opacity configurations that approach threshold instability; and (c) correction, because it guides targeted interventions (simplification, enhanced documentation, strengthened governance) where tolerance margins narrow.

6.6 TfO as a Systemic Variable (Preliminary Note)

Although the present paper remains confined to institutional calibration, it should be noted that widespread miscalibration of TfO across institutions may produce systemic effects. If multiple entities operate persistently beyond reconstructibility thresholds, supervisory review, capital comparability, and litigation stability may collectively deteriorate—this observation remains preliminary. The current framework addresses micro-prudential calibration; the macro-prudential implications of aggregate opacity accumulation—*Systemic Opacity Risk* (SOR)—constitute a distinct analytical layer.

6.7 Transition to Metric Formalization

The conceptual architecture so far: opacity degrades reconstructibility; reconstructibility anchors institutional accountability; TfO defines the maximum opacity compatible with stable reconstructibility. The next step is to translate this threshold logic into a structured calibration mechanism.

7. From Concept to Calibration

7.1 The Need for Operationalization

Supervisory practice requires structured methods of assessment, comparative reasoning, and threshold identification. Without such structuring, TfO would collapse into discretionary judgment or rhetorical caution.

The challenge, therefore, is not to transform *Tolerance for Opacity* into a rigid quantitative index, but to articulate a calibration architecture capable of preserving supervisory operability, avoiding artificial precision, accommodating institutional heterogeneity and capturing interaction effects among opacity domains.

Operationalization must remain proportionate. Over-formalization would create a false sense of numerical certainty; under-specification would render the concept

impracticable: TfO must therefore function as a structured prudential boundary condition rather than as a performance metric to be optimized[9].

The threshold is not presented as a predictive metric capable of determining collapse in advance. It functions as a prudential boundary concept. Its role is not to forecast failure, but to discipline institutional design by identifying configurations that materially increase fragility under supervisory or judicial stress.

7.2 Calibration as Institutional Practice

Calibration is not a purely technical exercise. The assessment of opacity levels and reconstructibility capacity necessarily involves structured judgment. It requires coordinated input from model developers, risk management functions, compliance units, internal audit, and senior management. In regulated banking environments, this internal calibration is subject to supervisory dialogue and review.

Reconstructibility presupposes the existence of an articulated AI governance architecture capable of allocating responsibilities, preserving documentation layers, and ensuring traceable decision pathways across institutional structures, such as proposed in *The Five Beacons Model*

Tolerance for Opacity is therefore not self-declared in isolation. It emerges from interaction between institutional self-assessment and supervisory expectations. This interaction resembles other prudential processes in which institutions articulate internal risk positions subject to external scrutiny. The distinctive feature of TfO lies in its focus on the structural sustainability of reconstructibility.

Calibration, in this context, performs three institutional functions: (a) mapping, what identifies where opacity accumulates across manyfold domains; (b) thresholding, which determines whether cumulative opacity approaches a level at which reconstructibility becomes unstable; and (c) governance, for it triggers corrective measures when tolerance margins narrow.

Crucially, calibration must remain iterative. AI systems evolve; retraining cycles alter internal architectures; outsourcing arrangements shift; data pipelines expand. A one-time assessment of opacity is insufficient. Tolerance for Opacity must be embedded within ongoing monitoring structures.

7.3 The Threshold Logic of Prudential Opacity

At the core of *Tolerance for Opacity* lies a threshold logic. Opacity does not become problematic merely because it exists, but because it may exceed the institution's capacity

[9] The emphasis on traceability and structured model governance aligns with supervisory expectations in prudential regulation. See Federal Reserve SR 11-7 (2011); BCBS 239 (2013); ECB Guide to Internal Models; EBA Discussion Paper on Machine Learning for IRB Models (2021).

to reconstruct AI-driven decisions under conditions of stress, dispute, or supervisory review.

Reconstructibility is not a static property; it is a capacity that must remain operational when most needed. The prudential concern is therefore anticipatory rather than reactive. The question is not whether opacity can be justified *ex post* in isolated instances, but whether aggregate opacity remains compatible with sustained institutional defensibility.

The threshold, accordingly, does not describe a moment of sudden systemic collapse. It marks the point at which opacity exposure begins to erode reconstructive reliability beyond an acceptable prudential margin. At that point, the difficulty is not merely technical degradation, but institutional incapacity to rearticulate decisional logic coherently under concentrated scrutiny.

TfO does not prohibit complexity. It defines the maximum opacity compatible with stable institutional reconstructibility under conditions of stress. Complexity remains admissible; opacity beyond the threshold does not.

Tolerance for Opacity, properly calibrated, functions as an early-warning boundary. Its effectiveness, however, presupposes governance arrangements capable of activating compensatory controls before reconstructive capacity deteriorates irreversibly.

7.4 From Threshold to Structured Calibration

Operationalizing TfO requires a structured assessment of the relationship between opacity exposure and reconstructibility capacity. This relationship can be expressed, at a structural level, through the condition:

$$O \leq RC / f(FC)$$

where opacity (O) must remain proportionate to institutional reconstructibility capacity (RC), modulated by the criticality of the function in which opacity operates (FC).

Functional Criticality is not determined solely by internal institutional assessment. In prudential contexts, its upper bounds are ultimately shaped by regulatory expectations and systemic relevance. The category therefore operates within an externally anchored normative environment.

This expression does not claim mathematical precision. It articulates a prudential discipline: AI opacity is admissible only insofar as it remains proportionate to institutional preparedness and to the risk sensitivity of the function concerned. In functions reaching extreme levels of systemic or prudential criticality, opacity may become normatively non-tolerable irrespective of institutional reconstructive strength. In such cases, the prudential constraint operates as a categorical boundary rather than as a variable threshold.

8. The Opacity–Reconstructibility Matrix

To visualize the proportionality without imposing artificial linearity, we introduce the *Opacity–Reconstructibility Calibration Matrix*. The Matrix does not replace judgment; it structures it. It offers a conceptual tool through which institutions and supervisors may assess whether opacity exposure remains within reconstructible bounds.

8.1 Structural Design of the Matrix

The operationalization of *Tolerance for Opacity* requires a structured representation capable of capturing the interaction between aggregate opacity and institutional reconstructibility capacity.

The *Opacity–Reconstructibility Matrix* is built on two axes: (i) Horizontal Axis: *Aggregate Institutional Opacity* (AIO), derived from the interaction of Model Opacity (MO), Process Opacity (PO), and Accountability Opacity (AO); (ii) Vertical Axis: *Institutional Reconstructibility Capacity* (IRC), reflecting the institution's ability to retrace, articulate, and justify decision pathways under formal scrutiny[10].

The matrix does not assume linear proportionality between the axes. Instead, it captures a dynamic relationship in which increases in opacity place progressively greater strain on reconstructibility mechanisms.

Aggregate AI opacity is not a simple arithmetic sum of MO, PO, and AO. It represents their combined institutional effect. High opacity in one domain may be partially offset by robustness in another; however, compensatory capacity is finite. Beyond certain configurations, reconstructibility begins to degrade non-linearly. The matrix therefore functions as a prudential map rather than a formula.

8.2 Aggregate Institutional Opacity (AIO)

Aggregate Institutional Opacity reflects the cumulative burden placed on reconstructibility mechanisms by structural model complexity (MO), operational retrievability constraints (PO) and fragmentation or dilution of responsibility (AO).

AIO is context-sensitive. For instance, a highly complex deep-learning model may generate substantial MO, yet remain prudentially manageable if PO is minimal and Accountability structures are robust. Conversely, moderate opacity across all three domains may collectively produce a level of aggregate opacity that exceeds institutional

[10] In order to avoid terminological inflation, the framework proposed herein relies on a conceptual hierarchy operating across five levels: (i) an ontological level, where *Opacity* resides as an inherent technical phenomenon; (ii) an institutional level, centered on *Reconstructibility* as the entity's capacity for accountability; (iii) a prudential level, where *TfO* serves as a manageable threshold; (iv) a modulating level, where variables such as *Functional Criticality* (FC**)** adjust said threshold; and (v) a systemic level, pointing toward future developments regarding *Systemic Opacity Risk (SOR***)**.

tolerance. AIO must therefore be assessed qualitatively and comparatively. It reflects the overall opacity profile of a specific AI deployment within its institutional setting.

8.3 Institutional Reconstructibility Capacity (IRC)

Institutional Reconstructibility Capacity measures the strength of the institution's compensatory architecture. IRC depends on factors such as quality and granularity of logging mechanisms, version control and traceability of model updates, internal technical expertise, independence and strength of validation functions, clarity of accountability chains or accessibility of third-party model components in outsourced arrangements.

IRC is not constant. It may improve through governance reinforcement or degrade through organizational complexity and resource constraints. Crucially, IRC does not eliminate opacity—it manages it. The vertical axis therefore reflects the institution's capacity to re-illuminate opaque decision processes when required[11].

8.4 Zones of Stability and Instability

When plotted against one another, AIO and IRC generate three prudential zones:

A, Stable Reconstructibility Zone

In this region, *Institutional Reconstructibility Capacity* comfortably exceeds the strain imposed by *Aggregate Institutional Opacity*. Characteristics include decision pathways can be retraced within reasonable timeframes; supervisory inquiries can be answered coherently; litigation defense remains evidentially robust; capital calculations can be justified without structural ambiguity. Opacity exists but remains institutionally contained.

B. Fragile Reconstructibility Zone

In this intermediate region, AIO approaches the limits of IRC. Compensatory mechanisms function, but with increasing strain. Early warning signals may include: extended time required to reconstruct specific decisions; dependence on a narrow group of technical specialists; difficulties in retrieving historical model states; supervisory queries requiring iterative clarification. This zone does not necessarily imply non-

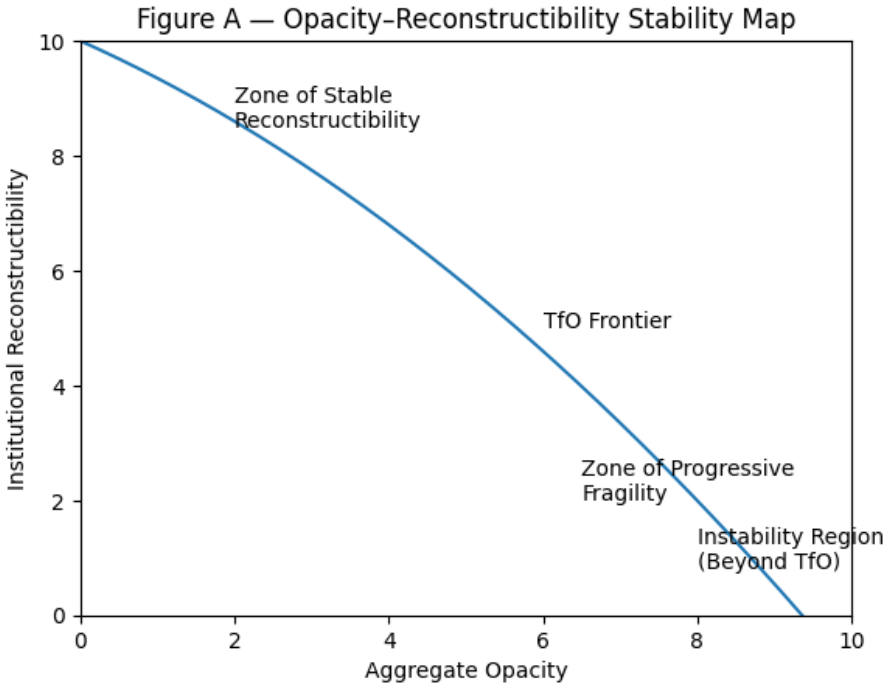
[11] The internal calibration of the IRC is not a purely abstract exercise. While this paper focuses on the prudential threshold (TfO), it is worth noting that the practical deployment of reconstructibility is significantly streamlined when an institution, for instance, adopts a governance architecture based on *The Five Beacons Model*. This multi-layered approach ensures that the data points required to populate the IRC are already structured and available, transforming the theoretical requirement of accountability into a systematic institutional habit. Thus, the Five Beacons framework serves as a reconstructibility enabler, allowing the institution to support higher levels of model complexity without breaching its *Tolerance for Opacity* (TfO).

compliance. However, it indicates that tolerance margins are narrowing. Incremental increases in opacity or minor governance failures may push the institution toward structural instability.

C. Impaired Reconstructibility Zone

In this region, *Aggregate Institutional Opacity* exceeds the institution's effective reconstructibility capacity. Indicators may include inability to reliably retrace specific decision pathways; fragmented accountability preventing coherent justification; substantive gaps in documentation or retrievable model states; material uncertainty in defending decisions under litigation or supervisory review. At this point, TfO has been exceeded. The institution is operating beyond its tolerance boundary. The impairment is structural rather than incidental.

Figure A



The Opacity–Reconstructibility Stability Map illustrates the gradual interaction between aggregate opacity and institutional reconstructive capacity. The curved frontier represents the Tolerance for Opacity (TfO) threshold: above it, opacity remains prudentially sustainable; below it, reconstructibility progressively deteriorates until institutional defensibility becomes unstable. The framework avoids binary classifications and instead emphasizes gradual degradation dynamics, consistent with the incremental nature of both opacity accumulation and governance erosion[12].

[12] An experimental alternative to this *TfO Governance Map* has been a quadrant-based visualization, where: (a) low-opacity/high-reconstructibility configurations correspond to traditional Glass-box models; (b) high-opacity/high-reconstructibility outlines define the *Stewardship Zone*, where technical opacity is offset by robust governance structures; (c) low-opacity/low-reconstructibility shapes might mean administrative

8.5 Non-Linearity and Threshold Effects

The relationship between AIO and IRC is non-linear. Reconstructibility does not degrade proportionally to opacity increases. Initially, increases in opacity may be absorbed by governance mechanisms with limited strain. However, once institutional buffers are saturated, additional opacity may produce disproportionately large reductions in reconstructibility capacity[13].

This threshold dynamic is critical. It implies that prudential assessment must focus not merely on current stability, but on proximity to inflection points. *Tolerance for Opacity* corresponds to the highest point along the opacity axis at which reconstructibility remains within the stable or manageable range. Beyond that point, the probability of abrupt impairment increases significantly[14].

8.6 Application to Banking Contexts

Within banking, the matrix has concrete prudential implications. In credit scoring models used for capital allocation, impaired reconstructibility may compromise the justification of risk-weighted assets. In anti-money laundering systems, opacity exceeding tolerance thresholds may undermine evidentiary defensibility. In automated customer decision systems, undermined reconstructibilty may weaken the institution's capacity to provide intelligible explanations to affected clients.

The matrix does not prescribe architectural simplicity. It permits advanced AI systems, provided that *Institutional Reconstructibility Capacity* evolves proportionately. Thus, the prudential imperative is symmetry: as opacity increases, reconstructibility capacity must augment correspondingly. Failure to maintain this symmetry results in tolerance breach.

8.7 The Matrix as Supervisory Instrument

The *Opacity–Reconstructibility Matrix* is not a scoring device but a structured evaluative framework. It enables internal mapping of opacity configurations; identification of tolerance margins; and supervisory dialogue grounded in reconstructibility rather than in abstract transparency demands. By shifting the focus from interpretability rhetoric to

---

negligence; and (d) high-opacity/low-reconstructibility figures represents a prudentially unacceptable composition in which institutional defensibility collapses. Main objection: the *Opacity–Reconstructibility Matrix* should not be interpreted as a set of rigid, compartmentalized quadrants, but rather as a dynamic field of forces where threshold graduality, contextual flexibility, and institutional dependencies determine the actual boundaries of supervisory acceptability.

[13] *Functional Criticality* refers to the prudential relevance of a given AI system within the institution's risk architecture. Systems directly affecting capital adequacy, liquidity stability, or legally significant decisions possess higher criticality. Opacity in high-criticality functions exerts disproportionate strain on *reconstructibility* thresholds.

[14] Interaction effects among opacity domains may operate as an *Opacity Risk Multiplier*, whereby moderate opacity in multiple domains generates disproportionate reconstructibility strain. This multiplier is context-dependent and not reducible to linear aggregation.

institutional reconstructibilty, the matrix anchors AI governance within the core prudential concern of accountability preservation.

9. Internal Calibration Mechanisms

9.1 Institutional Self-Assessment of Opacity Profiles

The *Opacity–Reconstructibility Matrix* provides a structural map. However, its prudential relevance depends on internal calibration mechanisms capable of translating that map into institutional practice. The maiden requirement is systematic self-assessment, which demands a robust architecture of AI governance, such the proposed *The Five Beacons Model*.

Institutions deploying AI systems in material decision functions should maintain a structured inventory of such systems, identifying those whose outputs affect capital allocation or risk-weighted assets; influence creditworthiness or pricing decisions; trigger regulatory reporting and/or produce legally relevant outcomes in relation to clients or counterparties.

For each material system, the institution should assess its opacity profile across the three domains (MO, PO, AO). This mapping exercise is not designed to produce a single numerical index. Rather, it produces a structured opacity profile, allowing institutions to visualize where cumulative strain on reconstructibility may arise.

Opacity assessment must remain proportionate: systems with marginal prudential relevance may warrant simplified evaluation; core risk engines and capital-relevant models require deeper scrutiny.

9.2 Reconstructibility Stress Testing

Opacity mapping alone is insufficient. Institutions must test reconstructibility under simulated scrutiny conditions. *Reconstructibility Stress Testing* (RST) is proposed as a procedural mechanism through which institutions evaluate their effective capacity to retrace decisions *ex post*.

An RST exercise may include: (i) selecting a historically issued decision (e.g., credit denial, risk classification, suspicious transaction alert); (ii) requiring the institution to reconstruct the complete AI-driven decision pathway using archived data, model versions, and governance documentation; (iii) assessing the time required to achieve reconstruction; (iv) evaluating the evidentiary coherence of the reconstructed explanation; and (v) identifying dependencies on specific individuals or external providers.

The objective is not to verify statistical correctness, but to test institutional retrievability under realistic constraints[15]. Stress testing may reveal latent fragilities: missing historical

[15] Stress testing and governance validation reflect established supervisory approaches to model risk management. See SR 11-7 (2011).

model states; insufficient documentation of retraining cycles; limited access to third-party model components; overreliance on informal expertise rather than structured documentation. Where reconstruction proves materially burdensome or incomplete, opacity tolerance margins may be narrower than initially assumed.

9.3 Governance Feedback and Corrective Measures

Calibration must be linked to corrective capacity. Where internal assessment or stress testing indicates that *Aggregate Institutional Opacity* approaches or exceeds tolerance thresholds, targeted governance adjustments become necessary—architectural simplification of specific models; enhanced logging and version control mechanisms; formalization of retraining documentation; strengthening of independent validation functions; clarification of accountability allocation in outsourced arrangements.

The objective is not to eliminate opacity but to restore symmetry between opacity accumulation and reconstructibility capacity. This feedback loop must operate continuously. Opacity is cumulative; governance erosion may occur gradually. Calibration therefore requires periodic reassessment rather than episodic review.

9.4 Opacity Accumulation Monitoring

A central vulnerability in AI governance is gradual opacity drift. As models are incrementally refined, retrained, or expanded in scope, complexity may accumulate without deliberate strategic choice: process documentation may degrade under operational pressure; organizational restructurings may fragment accountability chains.

Institutions should therefore monitor opacity accumulation longitudinally. This may involve periodic re-evaluation of opacity profiles; trigger thresholds linked to significant model updates; governance review following major outsourcing modifications; escalation protocols where reconstructibility indicators deteriorate. Opacity accumulation monitoring does not require precise quantification: it entails structured vigilance.

9.5 Supervisory Dialogue and External Review

Although TfO calibration begins internally, it does not remain purely internal. Supervisory review may legitimately focus not on the interpretability of specific model parameters, but on the robustness of the institution's reconstructibility architecture.

Supervisors may inquire: *How is opacity mapped? How frequently is reconstructibility stress-tested? What indicators signal proximity to tolerance thresholds? What corrective mechanisms are activated upon breach?* The presence of structured internal calibration mechanisms strengthens institutional defensibility. It demonstrates that opacity is managed deliberately rather than accumulated inadvertently.

9.6 Limits of Internal Calibration

Internal calibration cannot eliminate structural opacity. Nor can it substitute for ethical governance or fairness controls. Its purpose is narrower: to preserve the institutional capacity to retrace and justify AI-driven decisions under formal scrutiny.

Overconfidence in internal metrics may generate complacency. Formal compliance indicators should not replace substantive reconstructibility testing. Calibration must remain critical and adaptive. Ultimately, *Tolerance for Opacity* is sustained not by documentation volume, but by institutional discipline.

10. Supervisory Implications

10.1 Reframing Supervisory Focus

The Opacity–Reconstructibility framework does not require supervisors to engage directly with algorithmic code or technical architecture at a granular level; its relevance lies elsewhere. Supervisory attention shifts from the interpretability of individual models to the stability of institutional reconstructibility. The central question becomes: *Can the institution reliably retrace and justify AI-driven decisions under formal scrutiny?*

This reframing aligns AI governance with established prudential principles. Supervisors routinely assess capital adequacy, governance robustness, and risk management effectiveness without recalculating each internal model parameter. Similarly, the focus in AI deployment should rest on the resilience of the accountability structure rather than on exhaustive technical deconstruction.

10.2 Interaction with Existing Regulatory Frameworks

Tolerance for Opacity does not replace existing frameworks governing model risk, operational risk, or outsourcing arrangements. It operates at a different analytical layer. Model risk management addresses performance reliability and statistical validity; operational risk frameworks focus on process integrity and failure prevention; outsourcing regimes examine third-party dependencies and control retention.

TfO overlays these regimes by asking a distinct question: *Does the combined opacity profile of the institution compromise its ability to preserve reconstructibility?* In this sense, TfO is integrative rather than substitutive. It identifies structural conditions that may remain invisible when each risk domain is assessed in isolation.

10.3 Proportionality and Institutional Diversity

Not all institutions deploy AI at equivalent levels of complexity. Smaller entities may rely on externally supplied systems with limited internal technical capacity. Large cross-border institutions may operate highly sophisticated architectures supported by specialized validation units.

Supervisory application of TfO must therefore remain proportionate. The relevant inquiry is not absolute opacity, but opacity relative to institutional reconstructibility capacity. Advanced AI architecture may be prudentially sustainable within an institution possessing robust governance structures, while the same architecture may exceed tolerance thresholds in a less equipped environment. *Tolerance for Opacity* is thus inherently contextual.

10.4 Early-Warning Function

Supervisory engagement with TfO may serve an anticipatory function. Institutions operating persistently within the Fragile Reconstructibility Zone may not yet exhibit compliance failures. However, proximity to threshold instability increases vulnerability to litigation shocks, supervisory interventions, operational breakdowns, and governance fragmentation. By identifying narrowing tolerance margins, supervisors may intervene before structural impairment materializes—the objective is stabilization rather than sanction.

11. Prudential Safeguards Against Metric Illusion

11.1 The Risk of Over-Quantification (The Illusion of Precision)

Any attempt to operationalize *Opacity* invites numerical reduction. Scores, indices, and composite indicators promise clarity and comparability. Yet excessive quantification may obscure rather than illuminate the prudential problem. Opacity is multidimensional; reconstructibility is context-dependent. Interaction effects are non-linear. Reducing these dynamics to a single numerical score risks creating a metric illusion: apparent precision masking structural fragility. TfO should therefore resist rigid numerical fixation; calibration bands and structured qualitative assessment are preferable to artificial point estimates.

The objective is not to calculate AI opacity with mathematical exactitude. The objective is to detect structural configurations in which reconstructibility approaches impairment. Accordingly, our calibration architecture does not claim predictive precision. It provides a structured matrix through which institutions and supervisors may assess the stability of reconstructibility across opacity dimensions.

11.2 Gaming and Formal Compliance

A further vulnerability lies in strategic adaptation. Institutions may optimize documentation volume without enhancing substantive retrievability. Accountability structures may be formally clarified while practical responsibility remains diffused. Superficial simplifications may be introduced to improve opacity indicators without addressing deeper architectural strain. TfO must therefore be applied as a boundary condition, not as a performance target. It is not a score to be maximized or minimized: it is a limit to be respected.

11.3 Preserving Substantive Reconstructibility

The ultimate safeguard lies in periodic substantive testing. *Reconstructibility Stress Testing* must remain central. Formal compliance documentation cannot substitute for demonstrable ability to retrace concrete decision pathways. Where *ex post* reconstruction fails under realistic conditions, tolerance has been exceeded regardless of formal indicators. The integrity of the framework depends on preserving this substantive orientation.

12. Conclusion

Artificial intelligence expands the computational frontier of banking, but it does not alter the institutional architecture within which banking operates. Decisions affecting capital allocation, credit access, liquidity exposure, or compliance obligations remain subject to structured scrutiny. The legitimacy of AI-driven systems in regulated finance ultimately depends not on predictive sophistication, but on the preservation of institutional accountability.

Tolerance for Opacity should not be understood in isolation. Its normative coherence depends on broader institutional architectures that articulate explainability, oversight, and accountability functions in structured and mutually reinforcing layers. Without such architecture, tolerance thresholds risk becoming formalistic abstractions detached from operational reality.

This paper has proposed *reconstructibility* as the structural condition that secures that accountability. *Opacity* becomes prudentially relevant only insofar as it degrades the institution's capacity to retrace and justify decision pathways under supervisory, judicial, or fiduciary examination. *Tolerance for Opacity* (TfO) was defined as the maximum aggregate opacity compatible with stable reconstructibility.

The proposed *Opacity–Reconstructibility Matrix* does not attempt to quantify intelligence or optimize complexity. It introduces a calibration architecture through which institutions can preserve symmetry between algorithmic sophistication and accountability capacity. As opacity accumulates, reconstructibility must strengthen proportionately. Where this symmetry fails, institutional fragility emerges.

This framework does not seek to prohibit advanced AI architectures—nothing could be further from the truth. Nor does it equate transparency with simplicity. It does not eliminate complexity; it does not guarantee fairness, nor does it resolve broader ethical debates surrounding AI deployment. Its ambition is structural: to prevent reconstructibility material degradation within prudentially regulated environments.

Yet the implications may extend beyond immediate supervisory practice. If opacity accumulation were to become systemic—if institutions collectively operated near or beyond reconstructibility thresholds—the stability of supervisory coherence and risk-based capital regimes could be affected. In this sense, *Tolerance for Opacity* functions not

only as an internal governance tool but as a stabilizing principle within the broader architecture of regulated finance.

The automation of decision-making does not dissolve responsibility. It redistributes it across increasingly complex technical and organizational structures. Preserving reconstructibility ensures that responsibility remains institutionally anchored, even where computation exceeds human intuition. Opacity need not be eliminated, but it must remain governable. The durability of AI-driven decision-making depends not only on technical performance, but on sustained reconstructibility — and, in moments of crisis, on institutional forensibility.

The reconstructibility threshold framework is not intended to exhaust the interpretive space of prudential AI governance. It does not seek to replace supervisory judgment, institutional discretion, or judicial scrutiny. Rather, it proposes a structured vocabulary through which such judgment may be articulated, tested, and refined. Its parameters are necessarily provisional and invite calibration across institutions, jurisdictions, and sectors.

The framework proposed herein is intentionally structured to accommodate future refinements, including formal calibration models, sectoral adaptations, and cross-domain applications. Its present formulation should therefore be understood as a prudential scaffold rather than as a closed system.

The framework may appear conceptually demanding relative to current industry practices. However, prudential architecture is often constructed in advance of crisis recognition. TfO should therefore be read not as a description of prevailing standards, but as a proposal for structural resilience in anticipation of future stress scenarios.

If it contributes to the ongoing dialogue on algorithmic accountability in finance, it will do so not by closing interpretive possibilities, but by enabling them to be examined within a common architecture: a conceptual instrument offered for collective refinement in pursuit of systemic stability and the broader public interest.

Madrid, February 20th, 2026

Appendix A

Operational Calibration of the Reconstructibility Threshold

A.1 Preliminary Remarks

The analytical framework developed in this paper is intentionally structural and logic-driven rather than empirical, providing the structural parameters within which future quantitative implementation may occur. Nevertheless, the concept of *Tolerance for Opacity* (TfO) admits partial formalization. This Appendix does not propose a definitive quantitative model. It outlines possible avenues for structured calibration, intended to stimulate further analytical development. The following notes are illustrative.

A.2 Core Variables

For purposes of structured calibration, the following variables may be distinguished:

- MO (Model Opacity): architectural complexity and internal interpretability constraints.
- PO (Process Opacity): limitations in logging, version control, retrievability, and documentation.
- AO (Accountability Opacity): fragmentation or diffusion of responsibility within institutional or outsourced structures.
- IRC (Institutional Reconstructibility Capacity): the institution's effective capacity to retrace and articulate decision pathways under scrutiny.
- FC (Functional Criticality): the prudential importance of the AI system within the institution's risk architecture; e.g., capital relevance, systemic and litigation exposure.

A.3 Aggregate Institutional Opacity

Aggregate Institutional Opacity (AIO) may be expressed as a function of the three opacity domains:

$$AIO = f(MO, PO, AO)$$

The function f is not assumed to be linear. Interaction effects may amplify combined opacity beyond simple additive aggregation; for example, moderate opacity across all three domains may produce disproportionate reconstructibility strain compared to high opacity in a single domain.

Functional Criticality (FC) may operate as a contextual multiplier:

$$AIO_{adj} = AIO \times FC$$

Where FC reflects the prudential weight of the decision function in question. *Functional Criticality* is not infinitely compensable by reconstructive capacity. Beyond certain thresholds of systemic relevance, consumer impact, or capital sensitivity, opacity may become structurally intolerable irrespective of institutional preparedness. In such cases, the model does not yield a higher reconstructibility requirement; it establishes a ceiling on admissible

opacity. This reflects a fundamental prudential principle: some functions are too critical to be governed by compensatory logic alone.

A.4 Threshold Condition

The reconstructibility threshold condition may be expressed conceptually as:

$$IRC \geq AIO_{adj}$$

Reconstructibility remains stable where *Institutional Reconstructibility Capacity* equals or exceeds adjusted aggregate opacity. *Tolerance for Opacity* is exceeded when:

$$IRC < AIO_{adj}$$

This condition does not imply mechanical precision; it expresses a structural relationship: as opacity accumulates—particularly in high-criticality functions—reconstructibility capacity must increase proportionately.

A.5 Stress Coefficient and Dynamic Adjustment

Institutions may introduce a *Reconstructibility Stress Coefficient* (RSC), derived from stress testing exercises, to adjust IRC downward where reconstruction proves operationally fragile.

$$IRC_{eff} = IRC - RSC$$

Where RSC reflects empirical reconstruction difficulty observed in simulated scrutiny conditions. This adjustment introduces dynamic sensitivity to operational constraints.

A.6 Limits of Quantification

These expressions are illustrative abstractions. They do not provide universal coefficients, prescribed weightings, or regulatory scoring systems. *Opacity* and *reconstructibility* are context-sensitive and institution-dependent. Calibration requires structured judgment embedded within supervisory dialogue. The purpose of formalization is not to mechanize prudential reasoning, but to clarify structural relationships between opacity accumulation and accountability capacity.

A.7 Concluding Note

The reconstructibility threshold framework is capable of formal extension. Yet its normative force does not derive from mathematical expression. It derives from the institutional necessity of preserving accountability within increasingly opaque computational environments. Formalization may support that objective; it cannot substitute for it.

Bibliography / References

Ananny, Mike & Crawford, Kate. "Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society* 20, no. 3 (2018): 973–989.

Basel Committee on Banking Supervision (BCBS). *Principles for Effective Risk Data Aggregation and Risk Reporting (BCBS 239).* Basel, 2013.

Bovens, Mark. "Analysing and Assessing Accountability: A Conceptual Framework." *European Law Journal* 13, no. 4 (2007): 447–468.

Burrell, Jenna. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3, no. 1 (2016).

Doshi-Velez, Finale & Kim, Been. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv preprint (2017).

European Banking Authority (EBA). *Discussion Paper on Machine Learning for IRB Models.* EBA/DP/2021/02.

European Banking Authority (EBA). *Guidelines on Loan Origination and Monitoring.* EBA/GL/2020/06.

European Central Bank (ECB). *Guide to Internal Models.* Latest consolidated version.

Federal Reserve Board. *Supervisory Guidance on Model Risk Management (SR 11-7).* 2011.

Lipton, Zachary C. "The Mythos of Model Interpretability." *Queue* 16, no. 3 (2018).

Mittelstadt, Brent et al. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3, no. 2 (2016).

Rudin, Cynthia. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (2019): 206–215.

Wachter, Sandra, Mittelstadt, Brent & Floridi, Luciano. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7, no. 2 (2017): 76–99.

Zarsky, Tal. "Transparent Predictions." *University of Illinois Law Review* (2013): 1503–1570.

10.5281/zenodo.19144304