THE FIVE BEACONS MODEL

A Prudential Architecture for AI Explainability and Legal Liability in Banking

JM García-Maceiras

President of the Spanish BPO Banking Association

Title: *The Five Beacons Model: A Prudential Architecture for AI Explainability and Legal Liability in Banking*[1].

Abstract—The accelerated integration of Artificial Intelligence (AI) into core banking functions has generated a structural tension between algorithmic opacity and established prudential and liability frameworks. While horizontal governance standards—such as the National Institute of Standards and Technology AI Risk Management Framework and International Organization for Standardization ISO/IEC 42001—provide general principles of trustworthy AI, they do not fully address the evidentiary and supervisory demands specific to regulated financial institutions.

This paper advances the thesis that AI explainability deficits should be conceptualized as an autonomous prudential risk within banking supervision. The absence of structured explainability mechanisms may impair an institution's ability to demonstrate due diligence, sustain capital adequacy assessments, and withstand judicial scrutiny in complex liability scenarios.

To address this gap, the article introduces the Five Beacons Model (5B), a vertical, liability-oriented governance architecture designed to ensure decision-centric traceability across five institutional nodes: machine interface, engineering layer, human supervision, corporate governance, and client communication. Rather than treating explainability solely as a technical feature, the 5B model reframes it as an evidentiary infrastructure embedded within the Supervisory Review and Evaluation Process (SREP) and the Internal Capital Adequacy Assessment Process (ICAAP).

[1] A conceptual and prudential addendum to the Five Beacons Model as developed in the monograph *García-Maceiras, JM (2026). "The Banking Risk of AI Explanation: The Five Beacons Model". ADR Notebooks No. 1. Zyphrum Alchemists*. This governance architecture is the result of a multidisciplinary synthesis that integrates experience in the Judiciary and the senior banking leadership with a deep-rooted background in linguistic analysis and philosophical inquiry. By approaching Artificial Intelligence not merely as a technical challenge, but as a problem of legal semantics and causal logic, the author proposes a framework designed to restore intelligibility and accountability within complex financial systems.

By integrating prudential supervision with principles of tort law and causal attribution, the Model proposes a structured approach to mitigating opacity-related legal exposure and strengthening systemic resilience in AI-driven banking environments.

Keywords: *AI Governance, Banking Regulation, EU AI Act, Risk Management, Legal Causality, XAI, Financial Supervision, Five Beacons Model, SR 11-7, ICAAP*.

---

1. Introduction: The Explainability Gap in Global Banking

The global financial system is undergoing a paradigm shift. The integration of Artificial Intelligence (AI) into core banking operations—from credit underwriting and fraud detection to algorithmic trading—is no longer a competitive advantage, but a structural reality. However, this technological acceleration has outpaced the development of robust governance frameworks.

In this sense, explainability should be understood not only as a transparency mechanism but as an evidentiary safeguard embedded within governance architecture. By structuring documentation and oversight across institutional layers, the Model seeks to reduce fragmentation in causal reconstruction and to enhance the institution's defensive robustness in adversarial proceedings.

For the purposes of this paper, *explainability* is used as the umbrella concept encompassing *traceability* (technical reconstructibility of decision pathways), *intelligibility* (human comprehensibility across institutional levels), and *transparency* (normative disclosure obligations).

2. Methodological Positioning

This paper adopts a normative-operational methodology grounded in three analytical layers: (i) prudential banking supervision, (ii) tort law and causal attribution theory, and (iii) AI governance architecture.

First, the analysis draws upon the prudential framework developed under Basel III and European supervisory practice, particularly the Supervisory Review and Evaluation Process (SREP) and the Internal Capital Adequacy Assessment Process (ICAAP). These mechanisms provide the institutional setting in which governance deficiencies may translate into capital consequences.

Second, the article relies on principles of tort law, including doctrines of duty of care, burden of proof allocation, and evidentiary presumptions in complex harm scenarios. Rather than conducting a jurisdiction-specific doctrinal study, the paper adopts a comparative functional approach, focusing on recurring patterns in both civil law and common law systems concerning causality and negligence in technically complex environments.

Third, the 5B is presented as a governance architecture rather than a technical explainability method. It does not propose a specific XAI technique, but a structured allocation of traceability obligations across institutional nodes.

The objective is not to replace existing AI governance standards, but to articulate a sector-specific layer of accountability tailored to the evidentiary, supervisory, and capital implications of AI deployment in banking.

3. The Regulatory Imperative: Beyond the EU AI Act

The entry into force of the European Union AI Act marks the beginning of a new era of enforced transparency. For the banking sector, classified largely under high-risk use cases, transparency is increasingly embedded in binding regulatory obligations rather than remaining a purely ethical aspiration[2]. Notwithstanding, a profound gap exists between the high-level principles of the AI Act and the daily prudential reality of financial entities.

Current standards, while providing general guidance, fail to address the specific prudential density required by banking supervisors. The industry is currently trapped between technical methods that offer local interpretability (the *How*) and a regulatory environment that demands comprehensive accountability (the *Why*).

3.1. Explainability as a Prudential Risk

This proposal argues that the lack of AI explainability must be treated as an autonomous banking risk. It is not merely a subset of Operational Risk or Model Risk; it is a foundational risk that threatens: (a) Legal and Compliance Stability: the inability to justify a decision may expose the institution to supervisory findings and heightened litigation vulnerability; (b) Capital Adequacy: opacity in models can lead to higher capital charges as supervisors penalize unverifiable risk-weighted assets; and (c) Systemic Trust: in a sector built on confidence, high levels of model opacity may generate supervisory and evidentiary vulnerabilities in which errors can propagate without a traceable circuit breaker.

3.2. The Five Beacons Model (5B): A Sovereign Standard

To address this vacuum, this paper reaffirms the 5B. Developed through a multidisciplinary lens—combining technical AI governance, banking prudential standards, and a judicial understanding of causal liability—the 5B offers a vertical architecture for the financial sector.

The 5B do not merely seek to make AI interpretable for data scientists; they seek to make AI intelligible for the Board, auditable for the Supervisor, and defensible for the Judge.

[2] See *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*, particularly Arts. 9–15 (risk management, data governance, technical documentation, record-keeping, transparency, and human oversight requirements for high-risk AI systems). Financial services use cases such as credit scoring and risk assessment fall within the high-risk category under Annex III

By establishing five strategic nodes of information flow, the 5B ensures that the chain of responsibility remains unbroken, regardless of the complexity of the underlying algorithm.

4. The Judicial Nexus – Causality and Evidentiary Risk in AI-Based Decisions

The legal challenge posed by opaque AI systems is not primarily epistemological (how much a human can understand a machine) but evidentiary: how responsibility can be attributed when the decision-making process lacks traceable structure.

In complex tort litigation, courts frequently confront scenarios involving multiple actors and technical uncertainty[3]. Doctrines such as *res ipsa loquitur* in common law jurisdictions and evidentiary presumptions in civil law systems allow courts to mitigate informational asymmetry where the defendant exercises effective control over the relevant technical infrastructure[4].

In financial services, automated credit denial, transaction blocking, or algorithmic mispricing may generate similar evidentiary asymmetries. Where the institution cannot provide a coherent reconstruction of the decision pathway, it may face heightened litigation exposure, particularly in jurisdictions where courts insist on effective legal protection despite technological complexity[5]—not necessarily because opacity is unlawful *per se*, but because the inability to demonstrate due diligence weakens its defensive position[6].

The analogy to multi-vehicle collision litigation is instructive at a structural level. In such cases, judicial analysis focuses on reconstructing the chain of causation and identifying where the duty of care was breached. AI-driven decisions in banking similarly involve multiple layers: data sourcing, model design, deployment governance, and human oversight. Without documented traceability, the causal chain becomes fragmented.

---

[3] See *Restatement (Second) of Torts § 328D* (1965) (*res ipsa loquitur* doctrine, allowing inference of negligence where the instrumentality was under defendant's control and the harm would not ordinarily occur absent negligence). See also *Donoghue v Stevenson* [1932] AC 562 (HL) (foundational articulation of duty of care in negligence) and *Caparo Industries plc v Dickman* [1990] 2 AC 605 (HL) (foreseeability, proximity, and fairness criteria for duty of care). Civil law systems similarly recognize evidentiary presumptions and burden-shifting mechanisms in technically complex cases where information asymmetry disadvantages the claimant.

[4] See Court of Justice of the European Union, Case C-131/12, *Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González*, ECLI:EU:C:2014:317. The Court emphasized that entities exercising effective control over technologically mediated data processing remain subject to accountability obligations, notwithstanding the structural complexity of the underlying systems.

[5] See Bundesverfassungsgericht (German Federal Constitutional Court), *Judgment of 6 November 2019, 1 BvR 16/13* ("Right to be Forgotten I"). The Court confirmed that complex digital data-processing structures remain subject to constitutional standards of proportionality, transparency, and effective judicial protection, irrespective of their technical architecture.

[6] In situations characterized by informational asymmetry, courts may draw adverse inferences from the absence of documentation or traceability. See, e.g., *Restatement (Second) of Torts § 433B(3)* (burden of proof where multiple actors are involved). The inability to reconstruct a causal chain does not automatically establish liability, but it may materially weaken the defendant's evidentiary position.

The Model is designed to reduce this fragmentation. By requiring documentation and structured oversight at five institutional nodes, it seeks to provide *ex ante* evidentiary infrastructure capable of supporting *ex post* judicial scrutiny. Rather than presuming that opacity equates to negligence, the model acknowledges that irreversibly opaque systems may materially impair the institution's ability to discharge its burden of proof in adversarial proceedings.

In this sense, explainability functions not only as a transparency mechanism, but as a litigation-risk mitigation tool embedded in governance architecture. Transparency is the bridge that reconnects the act (the algorithmic output) with the consequence (the legal decision). The 5B operates as a structured attestation of the causal process.

Beyond the *broad brush* approach, it provides the judge with a traceable path from the initial data input to the final corporate oversight, effectively re-establishing the chain of command over the technology. The 5B seeks to reduce evidentiary uncertainty by providing a structured reconstruction of the decision-making chain—*precision as defense*.

5. The Five Beacons Model Architecture

The 5B is not a mere set of ethical guidelines; it is a multilayered governance architecture designed to ensure that AI-driven decisions are intelligible, traceable, and legally defensible. Each *Beacon* represents a critical node where the flow of information must be captured to maintain the chain of causality.

5.1. The Machine-to-Machine (M2M) Interface – Technical Traceability

At the foundational level, this Beacon addresses the raw data and algorithmic lineage. It moves beyond simple logging to establish a *Black-Box recorder* for AI. In high-frequency trading or automated credit scoring, it provides the granular evidence required for post-hoc forensic analysis. It is the technical anchor of the explainability chain; automatic generation of metadata regarding data inputs, weights, and versioning.

5.2. The Engineer-to-Model (E2M) Interface – Interpretability

This Beacon focuses on the translation layer. It is the space where data scientists and AI engineers must translate mathematical outputs into human-understandable features; implementation of local and global explanation methods (e.g., feature attribution) that are not just technically accurate but contextually relevant to the banking domain. The 5B ensures that the *Why* of a model's output is documented by the developer, preventing the drift of accountability from the creator to the user.

5.3. The Human-in-the-Loop (HITL) Supervisor – Real-Time Control

This is the most critical beacon for regulatory compliance. It establishes the protocol for the human supervisor who oversees the AI's output—both internal (corporation) and external (institutional supervision). The judicial angle of such transition is that, based on the principle of *duty of care*, the 5B provides evidence that the bank maintained effective

control over the automated process—not just a passive viewer, but an active gatekeeper with the authority and tools to override the system. It solves the automation bias problem by documenting human intervention (or the lack thereof).

5.4. The Corporate Governance Beacon – Board Oversight

This Beacon elevates AI risk to the highest level of the institution. It integrates AI explainability into the Risk Appetite Framework (RAF) and the Internal Capital Adequacy Assessment Process (ICAAP). The Board of Directors must receive structured reports on the intelligibility status of the bank's AI portfolio. In banking practice, it strengthens the Board's ability to demonstrate effective oversight in the event of supervisory or judicial review by demonstrating a structural commitment to AI transparency, moving AI from the IT department to the Corporate Governance agenda.

5.5. Fidelity and Transparency Node: Consumer Protection as Systemic Resilience

The final beacon closes the loop by addressing the end-user. It focuses on the *Right to Explanation* enshrined in regulations like GDPR and the EU AI Act[7]. Providing the client with a clear, concise, and non-technical justification for a decision (e.g., a loan denial), this beacon addresses informational asymmetry and reduces litigation exposure. In banking law, courts have increasingly required not only formal disclosure but effective intelligibility for clients in structurally imbalanced relationships[8]. A client who understands *why* is less likely to challenge the decision in court. It transforms transparency into a competitive advantage and a trust-building tool.

6. Comparative Positioning of the Five Beacons Model

The current landscape of AI governance is dominated by horizontal standards that, while valuable for general purposes, fail to address the idiosyncratic risks of the banking sector. Standards such as the NIST AI Risk Management Framework (RMF) and ISO/IEC 42001 offer a one-size-fits-all approach that proves insufficient when confronted with the rigors of financial supervision and the precision required in judicial proceedings.

6.1. Beyond NIST AI RMF: From Voluntary Guidance to Mandatory Accountability

The NIST AI RMF is a benchmark for flexibility and voluntary adoption. However, for a Tier-1 financial institution, *flexibility* is often a synonym for *legal uncertainty*. NIST

[7] See *Regulation (EU) 2016/679 (General Data Protection Regulation)*, Art. 22 (automated individual decision-making), Arts. 13–15 (transparency obligations), and Art. 82 (right to compensation for material or non-material damage). Although the existence and scope of a standalone "right to explanation" remain debated in academic literature, the GDPR clearly establishes transparency and accountability obligations in automated decision contexts.

[8] See Tribunal Supremo (Spain), Civil Chamber, *Judgment No. 241/2013, 9 May 2013* (*Cláusulas Suelo*). The Court developed the doctrine of "material transparency" in banking contracts, requiring that contractual terms be not only formally disclosed but sufficiently intelligible to the average consumer in order to ensure effective consent in asymmetrical financial relationships.

focuses on broad characteristics of trustworthiness (bias, safety, resilience). It tells organizations *what* to measure but remains silent on *how* to link those measurements to a chain of legal responsibility.

While NIST is an engineering-centric framework, the Model is liability-centric. It transforms NIST's abstract trustworthiness into a series of enforcement nodes. In a sector where every decision can lead to a systemic impact, the voluntary orientation of NIST is complemented by a structurally embedded allocation of responsibility more suited to regulated financial institutions.

6.2. Beyond ISO/IEC 42001: The Fallacy of Procedural Compliance

ISO 42001 provides a Management System (AIMS) that focuses on processes. Yet, in the eyes of a regulator or a judge, a certified process does not equate to a justified outcome. One can be *ISO-certified* and still operate a *Black-Box* model that violates the principle of causality. ISO 42001 primarily addresses governance processes, whereas the 5B focuses on the evidentiary substance of decision-making pathways.

Drawing from a deep judicial understanding of tort law and the nexus of causality, the 5B Model recognizes that a *procedural seal* is a weak defense in litigation. From a judicial perspective, courts do not primarily look for certified processes, but for traceable decision-making pathways. The 5B Model ensures that the explanation is not a *post-hoc* documentation exercise, but a structural guarantee of accountability that can withstand the scrutiny of an adversarial trial.

6.3. The Evolution of SR 11-7: AI-Native Model Risk Management

The Federal Reserve's SR 11-7 has long been the gold standard for Model Risk Management (MRM)[9]. However, it was designed for traditional econometric models, not for the stochastic and opaque nature of Generative AI or Deep Learning.

The 5B Model acts as the necessary evolution of MRM. It takes the principles of internal validation and governance from SR 11-7 and adapts them to the *explainability crisis* of AI. It moves from *model-centric* validation to *decision-centric* traceability.

7. Prudential Implementation: SREP, ICAAP, and the Supervisory Dialogue

The true test of any banking governance framework is its ability to be integrated into the existing prudential architecture. The Model is designed to function as an operational plug-in for the Supervisory Review and Evaluation Process (SREP) and the Internal Capital Adequacy Assessment Process (ICAAP).

[9] Board of Governors of the Federal Reserve System, *SR 11-7, Supervisory Guidance on Model Risk Management* (April 4, 2011). The guidance emphasizes model validation, governance, and documentation, though it was developed primarily with traditional statistical models in mind.

7.1. Integration into the Risk Appetite Framework (RAF)

A bank's appetite for AI risk cannot be measured solely by the potential for financial loss. It must include a *Tolerance for Opacity*. By adopting the 5B, an institution can set quantitative and qualitative thresholds for AI explainability. If a model fails to satisfy the requirements of Beacon 2 (Engineer) or Beacon 3 (Supervisor), it should automatically trigger a Risk Limit Breach, requiring immediate mitigation or the disconnection of the system.

7.2. *Tolerance for Opacity* (TfO) as a Prudential Metric

Traditional risk appetite frameworks quantify exposure in financial or operational terms. However, AI deployment introduces a distinct variable: the degree of acceptable decision opacity.

Tolerance for Opacity (TfO), as proposed in this paper, may be defined as the maximum level of algorithmic non-traceability that an institution is willing to assume without compromising its ability to: (a) demonstrate compliance with supervisory expectations; (b) reconstruct causal chains in litigation; (c) justify risk-weighted asset calculations; and (d) provide intelligible explanations to affected clients.

The determination of the TfO must not be viewed as a static threshold, but as a dynamic risk-weighted parameter integrated into the bank's internal capital assessment. A high *Tolerance for Opacity* in non-critical administrative processes may be permissible; however, in systemic functions such as credit underwriting or liquidity stress-testing, an elevated TfO acts as a direct multiplier of operational risk. By quantifying the intelligibility gap of a model, the institution can proactively calibrate its Pillar 2 capital add-ons, transforming an abstract algorithmic opacity into a manageable financial buffer. This ensures that the bank’s capital position remains resilient even when the underlying technology defies traditional forensic reconstruction.

From a liability perspective, the formal adoption of a TfO framework serves as an essential instrument of *ex-ante* due diligence. In the event of litigation or regulatory enforcement, the ability to demonstrate that a specific level of opacity was not an overlooked defect, but a consciously managed and mitigated strategic choice, may influence the allocation and practical dynamics of the burden of proof. By documenting the technical and human redundancy nodes that compensate for a model's opacity, the bank effectively re-establishes the chain of causality. Thus, a well-defined *Tolerance for Opacity* functions as a structural defense, shifting the judicial focus from the inherent inscrutability of the machine to the demonstrable robustness of institutional oversight.

Operationalization of this tolerance may rely on a combination of quantitative (i) and (ii) qualitative indicators, including: (i) percentage of AI portfolio covered by documented explainability protocols; override frequency and review documentation rates, and time required to reconstruct decision logic; and (ii) a board-level reporting on explainability

gaps; escalation triggers when models exceed predefined opacity thresholds; and integration of explainability metrics into ICAAP stress scenarios[10].

From a prudential perspective, excessive opacity may generate supervisory concern analogous to deficiencies in internal controls. While current regulation does not formally define opacity thresholds, supervisors may interpret governance opacity as a qualitative weakness under Pillar 2 assessments. The 5B provides a structural method to measure and manage this tolerance, converting opacity from an abstract technological feature into a governable prudential parameter.

7.3. Impact on Pillar 2 Requirements (P2R)

Supervisors (e.g. ECB, EBA) increasingly penalize governance deficiencies with additional capital requirements under Pillar 2[11]. An opaque AI system represents an unquantified operational risk. The Model serves as a mitigation technique. By proving that the chain of causality is documented through the five nodes, the bank can argue for a lower risk profile, potentially reducing the capital add-ons that would otherwise be imposed due to algorithmic uncertainty.

7.4. The Supervisory Interface: Streamlining the Dialogue

One of the greatest challenges for banks is responding to *Section 17 style* inquiries or on-site inspections regarding automated decisions. The 5B Model provides a standardized handbook, a common language for both the bank and the inspector. Instead of *ad-hoc* explanations, the bank presents a *Beacons Report*. This structured documentation proves that the institution has moved from *reactive compliance* to *proactive architectural governance*, significantly reducing the duration and friction of supervisory reviews.

8. Conclusion – Towards a Global Standard of Accountability

The increasing deployment of opaque systems raises significant supervisory and legal concerns—not because technology demands it, but because the law requires it. As this paper has demonstrated, the challenge of AI explainability is fundamentally a judicial and prudential one.

The Model offers a sector-specific governance alternative. By shifting the focus from technical interpretability to architectural accountability, the 5B: (i) reduces the risk of adverse burden-of-proof dynamics in litigation; (ii) protects the Board by providing the necessary eyes to exercise their duty of oversight; and (iii) protects the Financial System

[10] The operational categorization of TfO thresholds across specific banking clusters—including the development of a formal *TfO Calibration Matrix* for internal capital adequacy assessments—will be the subject of forthcoming research papers by the author, expanding upon the architectural foundations established herein.

[11] See European Banking Authority (EBA), *Guidelines on the revised common procedures and methodologies for the Supervisory Review and Evaluation Process* (SREP); European Central Bank (ECB), *Guide to the Internal Capital Adequacy Assessment Process* (ICAAP); and *Basel Committee on Banking Supervision, Core Principles for Effective Banking Supervision.* Governance weaknesses and deficiencies in internal control frameworks may result in qualitative measures and additional capital requirements under Pillar 2.

by ensuring that every automated decision remains anchored to a human and legal responsible party.

In an increasingly automated financial environment, institutional resilience will depend less on algorithmic sophistication alone and more on the robustness of governance structures surrounding automated decision-making. The Five Beacons Model proposes one possible architecture for aligning technological deployment with supervisory accountability and legal defensibility.

Madrid, January 25th, 2026

References / Bibliography

Basel Committee on Banking Supervision (BCBS). *Newsletter on Artificial Intelligence and Machine Learning in Banking* (2024/2025 updates) and *Core Principles for Effective Banking Supervision*.

Board of Governors of the Federal Reserve System / Office of the Comptroller of the Currency (OCC). *SR Letter 11-7: Guidance on Model Risk Management* (April 4, 2011).

European Banking Authority (EBA). *Guidelines on the revised common procedures and methodologies for the supervisory review and evaluation process (SREP) and supervisory stress testing.*

European Banking Authority (EBA). *Guidelines on Internal Governance under Directive 2013/36/EU* (EBA/GL/2021/05).

European Banking Authority (EBA). *Guidelines on ICT and security risk management* (EBA/GL/2019/04).

European Central Bank (ECB). *Guide on Model Risk* (March 2024).

European Central Bank (ECB). *Guide to the Internal Capital Adequacy Assessment Process (ICAAP)* (November 2018, updated expectations 2024).

European Central Bank (ECB). *The Supervisory Review and Evaluation Process (SREP): Aggregated results and expectations.* (Annual Reports 2023-2025).

European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).*

International Organization for Standardization (ISO). *ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system.*

National Institute of Standards and Technology (NIST). *AI Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce (2023). DOI 10.5281/zenodo.18647317