Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Consultation report

Response to Consultation

Asociación Española de Externalización de Procesos y Servicios (AEPROSER)

1. **Do you agree with the benefits and risks of AI adoption by financial institutions described in this report? Are there any substantive benefits or risks not covered?**

   I broadly agree with the benefits and risks identified in the report. The document appropriately highlights opportunities related to efficiency, productivity, customer service, fraud prevention, and risk management, while also addressing governance, operational, model, third-party, and cyber-related risks.

   One additional area that may deserve further consideration is the governance challenge created by increasing levels of system opacity. Even where AI systems perform satisfactorily and remain formally compliant, financial institutions may encounter difficulties in reconstructing the chain linking data, models, decisions, oversight actions, and accountable decision-makers when outcomes are challenged by customers, supervisors, auditors, or courts.

   This exposure may not necessarily arise from model failure, but from limitations in the institution's ability to demonstrate how a specific outcome was produced and under whose responsibility it ultimately falls. As AI systems become more complex, this issue may affect not only individual institutions but also confidence in AI-enabled decision-making within the financial system. Maintaining the capacity to reconstruct decisions under scrutiny may therefore become an increasingly important component of responsible AI governance.

2. **Are the sound practices sufficiently comprehensive and clear to enable financial institutions' responsible AI adoption?**

   The sound practices are comprehensive and provide a useful governance-oriented framework for responsible AI adoption.

   One area that could be further emphasized is the distinction between implementing governance controls and preserving the institution's ability to sustain accountability when AI-enabled decisions are challenged. Alongside explainability, accountability and traceability, financial institutions may benefit from ensuring that decisions remain capable of being reconstructed throughout the AI lifecycle.

In practical terms, this requires governance arrangements that preserve the evidentiary links between data, models, outputs, human oversight actions and accountable decision-makers. Such capabilities may become increasingly important as AI systems grow in complexity and as institutions are required to demonstrate the basis of specific decisions to supervisors, auditors, courts or customers.

3. **Do the sound practices strike an appropriate balance between managing risks relating to all forms of AI, and addressing some of the risks relating to emerging and new complex forms of AI, such as GenAI and agentic AI?**

   The report generally strikes an appropriate balance between addressing AI-related risks broadly and recognizing the additional challenges posed by emerging forms of AI, including GenAI and agentic systems.

   At the same time, governance frameworks may benefit from focusing less on specific AI architectures and more on the degree of opacity that institutions are capable of governing effectively. While technologies will continue to evolve, the underlying governance challenge often remains the same: whether institutions can understand, monitor, validate, oversee and remain accountable for system behavior at an acceptable level.

   For this reason, a risk-based approach centered on the institution's capacity to manage increasing levels of opacity may provide a more durable governance principle than one tied to particular technologies. Such an approach could remain relevant across successive generations of AI systems, including those that have yet to emerge.

4. **Are the sound practices sufficiently flexible to accommodate and address newer types of AI and responsible adoption over time?**

   Yes. The principles-based nature of the sound practices should allow them to remain relevant as AI technologies continue to evolve.

   Their long-term effectiveness, however, is likely to depend on maintaining a focus on governance outcomes rather than on specific technological characteristics. Principles such as accountability, oversight, traceability, validation and risk ownership are expected to remain relevant regardless of future developments in AI capabilities.

   In this respect, governance frameworks may be most resilient when anchored in enduring institutional objectives rather than in particular AI models, architectures or technological generations.

5. **Do the case studies in this report sufficiently highlight how different types of financial institutions can benefit from responsible AI adoption? Are there additional case studies for inclusion in the report? If so, please provide such case studies, particularly for nonbanks.**

   The case studies are useful and demonstrate practical applications of responsible AI adoption across a range of operational contexts.

   An additional case study that may be valuable would involve a disputed AI-enabled credit, underwriting, fraud detection, or customer risk classification decision. Such a scenario could

illustrate how governance, documentation, traceability, human oversight and accountability mechanisms operate when an institution is required to explain, evidence and defend a specific outcome before supervisors, auditors, courts or customers.

This type of case may be particularly relevant for both banks and nonbank financial institutions, as it highlights how responsible AI governance functions not only during model development and deployment, but also when decisions are subject to external scrutiny and challenge.

6. Do the case studies in this report provide actionable insights for financial institutions in their responsible AI adoption?

Yes. The case studies provide practical examples that help translate the sound practices into operational and governance contexts.

Their value lies not only in illustrating AI adoption opportunities, but also in demonstrating how governance, oversight, risk management and accountability mechanisms can be applied throughout the AI lifecycle. This practical dimension may be particularly useful for institutions seeking to operationalize principles-based guidance.

As AI use cases continue to evolve, additional case studies involving situations of external scrutiny, contested outcomes or governance escalation may further strengthen the report's practical relevance.

7. Are the definitions in the glossary clear and aligned with industry sound practices, including recent developments in AI?

The glossary is generally clear and consistent with current industry practice.

One possible enhancement would be to distinguish more explicitly between explainability and the broader institutional capacity to reconstruct AI-enabled decisions when they are subject to scrutiny. Explainability is typically concerned with understanding model behavior, outputs or decision logic. By contrast, what may be described as "reconstructibility" concerns the institution's ability to recover the evidentiary and governance chain linking data, models, decisions, oversight actions and accountable decision-makers.

While related, these concepts address different dimensions of responsible AI adoption. As AI systems become increasingly complex, the capacity to reconstruct how decisions were produced and governed may become an important complement to explainability, particularly in supervisory, audit, dispute resolution and accountability contexts.

8. Are there any comments you would like to make on this report? Please fill in.

The report provides a valuable contribution to the ongoing discussion regarding responsible AI adoption in financial services.

One broader observation is that, as AI systems become increasingly integrated into high-impact financial decision-making processes, governance challenges may progressively shift from questions of model performance alone toward questions of institutional accountability, evidentiary robustness and the capacity to sustain decision ownership under scrutiny.

In this context, future discussions may benefit from continued attention not only to how AI systems perform, but also to how institutions preserve effective governance, oversight and responsibility throughout the lifecycle of AI-enabled decisions.

Additional supporting materials are provided separately by email for reference.

BANKING & ARTIFICIAL INTELLIGENCE

A Research Series on AI Governance and Institutional Control

JM García-Maceiras

MONOGRAPH

The Banking Risk of AI Explanation

ARCHITECTURAL EXTENSIONS

Developing, layer by layer, a structured architecture for AI accountability in banking

The Five Beacons Model (5B)

Tolerance for Opacity (TfO)

Systemic Opacity Risk (SOR)

The HRAIS Chamber

Compiled Edition

Submitted as supporting material for ongoing policy and supervisory discussions

MONOGRAPH

THE BANKING RISK OF AI EXPLANATION

Foundational framework on AI explainability as a prudential and legal risk in banking

JM García-Maceiras

ARTIFICIAL INTELLIGENCE

THE BANKING RISK OF AI EXPLANATION

The Five Beacons Model

JM García-Maceiras

ZYPHRUM ADR NOTEBOOKS Nº1

JM García-Maceiras

THE BANKING RISK OF AI EXPLANATION

Front Cover: *AI Bank*
(*Author's AI Agents*)

The Banking Risk of AI Explanation
JM García-Maceiras

1st English Edition: January 2026

Zyphrum Alchemists

ISBN 9789403845760
Printed in the European Union

For information, please address presidencia@aeproser.com

Table of Contents

Title: The Banking Risk of AI Explanation

Abstract: Artificial Intelligence brings significant benefits to banking, but also introduces novel risks, such as the requirement that decisions made by complex machine learning systems and deep neural networks be adequately explained to humans. This work offers a multifaceted view of the topic, harmonizing the legal frameworks of Data Protection and Artificial Intelligence, the background of systems engineers, academic contributions from the computing community, and banking risk management under Basel III and the DORA Regulation, suggesting the idea of the Five Beacons *as a structural model for explanation to strengthen the protection of financial customers (and citizens at large) against the machine.*

Keywords: Artificial Intelligence; Explainability; Banking; Risk.

I

BANKING AND ARTIFICIAL INTELLIGENCE

Aligned with its essential mission of gathering the most reliable information to make optimal decisions, the banking industry has consistently embraced breakthrough technologies, provided that they have not induced new risks or undermined well-established methodologies and processes.

The history of financial institutions is, in many aspects, the story of how technological innovation has enhanced an economic activity of paramount importance to the progress of nations[1]. Among the milestones are the Italian invention of the bill of exchange in the 12th century, with its remarkable abstraction of a complex legal transaction; the advent of double-entry bookkeeping during the Renaissance; and the establishment of branch and correspondent networks in the 17th century, recognized today as early seeds of globalization.

Mechanization in the 18th and 19th centuries brought ingenious counting and adding machines, such as the arithmometer based on Leibniz's wheel. Later, the widespread adoption of the telephone and telegraph introduced a development that radically transformed communication: the immediacy of the message. Before the emergence of fax, email, and instant messaging, information on a global scale—including financial market news—was transmitted through a network of teletype machines.

[1] Ferguson, Niall (2009). "*The Ascent of Money: A Financial History of the World"*. Penguin Books.

With the dawn of the Information Age came the installation of the first computers (*mainframes*), admittedly of ungainly appearance; the automated teller machines; internal and operational digitization through online systems; the mass adoption of credit cards; and the installation of point-of-sale terminals in shops[2].

Over the past decade, the banking sector has undergone a profound digital transformation that has improved operational efficiency and customer experience. Alongside this modernization, complete with its advantages and drawbacks, other technologies have emerged to facilitate integrated information management and more sophisticated risk control.

Disruptive ideas such as Big Data, Biometrics, Cloud Computing, Smart Contracts, digital wallets; as well as Distributed Ledger Technology—primarily associated with blockchain and its derivatives, such as cryptoassets and tokens—together with emerging fields such as Quantum Computing and the universal Interconnection of Objects, are becoming decisive in reshaping the financial system and redefining banking activity.

The AI Bank

As we move through the first quarter of the 21st century, it can be asserted that, among these, none will have as far-reaching or enduring an impact on banking as the breakthrough now underway: the combination of Artificial Intelligence, Data Science, and Robotics into a single transformative framework (ADR)[3].

Artificial Intelligence (AI; *Artifiligence*[4]) means *a machine-based system that is designed to operate with varying levels of*

[2] Walker, Tim; Lucian Morris (2021). *"The Handbook of Banking Technology".* Wiley.

[3] ADRA-The AI, Data, Robotics Association (2023). "*Strategic Orientation towards an AI, Data, Robotics roadmap 2025-2027*" (May 2023). https://adra-e.eu/publications#

[4] For the sake of broadening the rather scanty linguistic stock on the topic, let's introduce a portmanteau.

autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments[5].

The reasons why AI has captured the attention of financial institutions are numerous and diverse, driven by the pursuit of higher productivity, lower costs, and greater quality and security in banking operations[6].

Properly applied, AI can optimize internal business processes, reduce costs, automate routine tasks, repurpose human resources toward higher-value activities, and increase productivity. It can enhance ICT applications by generating code from natural language, detecting and correcting programming errors, converting code between languages, and facilitating legacy system migration.

AI strengthens fraud detection by identifying anomalies in transaction patterns—amounts, frequencies, counterparties—and issuing alerts for further investigation. It refines risk modeling and management, improves anomaly detection, and supports better anticipation of market movements and customer behavior shifts.

AI's predictive capabilities enhance investment analysis, enabling more accurate and consistent forecasts in volatile markets, leveraging data aggregation from previously inaccessible sources and real-time updates.

---

[5] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts*. *Official Journal of the European Union*, L 2024/1689, 14 June 2024. / Also: *European Commission (2025). "Guidelines on the definition of an artificial intelligent system established by Regulation (EU) 2024/1689 (AI Act)".*

[6] Aldasoro, Iñaki; Et al. (2024). "*Intelligent financial system: how AI is transforming finance"* BIS Working Papers, 1194, Bank for International Settlements. https://www.bis.org/publ/work1194.pdf

In customer service, AI improves chatbot interactions, assists with internal employee inquiries, and automates transcription, summarization, and evaluation of call-center communications. It streamlines document itemization and meeting minute preparation.

For Legal Departments, AI aids in monitoring regulatory changes, synthesizing case law and academic commentary, analyzing contractual clauses, and suggesting revisions. It can process massive unstructured datasets, uncovering non-obvious patterns that refine client profiling and transaction clustering.

AI enhances creditworthiness assessments, improves customer verification (including remote onboarding and digital identification), and strengthens anti-money-laundering and counter-terrorist-financing monitoring. It supports real-time transaction analysis, improves compliance, and raises overall data quality, granularity, and precision.

Ultimately, AI contributes to profitability, customer satisfaction, and financial inclusion, expanding banking accessibility while ensuring greater accuracy, transparency, and regulatory diligence[7].

Survey Results: Perceptions of GPAI in Banking

According to a confidential survey conducted during 2025 by our employers' association[8] among banking professionals at all hierarchical levels, the vast majority reported that General-Purpose Artificial Intelligence (GPAI) is expected to yield significant benefits across multiple dimensions of banking activity. These include customer retention, proactive identification of client needs, optimization of pricing and commissions, talent acquisition and retention, streamlining of omnichannel relationships, enhanced

[7] Riemer, Stende; Et. al. (2023). "*A Generative AI Roadmap for Financial Institutions*". Boston Consulting Group. https://www.bcg.com/publications/2023/a-genai-roadmap-for-fis

[8] Spanish Association of BPO (AEPROSER) (2026). https://www.aeproser.com/the-fringe

reporting and summarization, automated client statements, transaction monitoring, cross-selling, and customer acquisition.

Respondents also highlighted the potential of GPAI to improve the automatic classification and labeling of documents, database agent training, contract monitoring, and the management of general clauses and standardized forms. Other advantages cited include improved market trend analysis, automated reporting of suspicious activities, intelligent routing, sales-force training, continuous due diligence, differentiated collections, optimization of commercial networks, code generation, and document completion.

Furthermore, activities identified as significantly enhanced by GPAI encompass error correction across all functional areas, intelligent document processing and digitization, early-warning generation, hyper-personalization of texts and images, new client onboarding, collateral risk assessment, call transcription and analysis, automated loan approval, risk-weighted asset optimization, offer personalization, and long-term strategic planning.

Even infrastructure management and auxiliary services, such as security, catering, and childcare, are perceived as potential beneficiaries of GPAI. The only domain where no substantial advantages were reported is the preservation of current employment levels for human staff within the sector.

A Typical Use Case: Credit Scoring

Within a relatively short time, GPAI has demonstrated its transformative potential in one of the banking sector's most emblematic processes: credit scoring.

GPAI enhances information processing by replacing not only manual data entry and validation but also technologies such as *Optical Character Recognition* (OCR). Leveraging Machine Learning (ML) algorithms and *Intelligent Document Processing* (IDP), it can interpret the semantic context of documents and extract

meaningful, structured information from unstructured formats. This reduces operational workload, improves accuracy, and minimizes costly errors and the risk of financial penalties. By digitizing and validating information at the source, GPAI creates clean, structured databases that effectively support subsequent algorithms and analytical models.

Traditionally, risk analysis has relied on structured data to predict default probability. Classical statistical techniques—such as logistic regression or decision trees—have been used to process financial and sociodemographic variables, as well as key ratios like *loan-to-value*. However, these models often fail to capture complex relationships between variables or integrate unstructured data.

GPAI is reshaping this analytical landscape. Deep Neural Networks (DNNs) and advanced ML techniques—such as *Random Forests*, *Support Vector Machines* (SVMs), *XGBoost*, and *Deep Learning* architectures—combined with *Natural Language Processing* (NLP) for analyzing customer comments and communications, offer far greater predictive accuracy than traditional methods. These algorithms detect complex, non-linear patterns within massive datasets, forecasting default rates with higher precision and identifying potential credit risks before they materialize[9].

The benefits are evident: decisions are reached in minutes rather than days; false negatives decrease from approximately 12% to 5%; default rates over 12 months fall by an average of 20%; and compliance with regulatory transparency requirements improves significantly.

However, the primary obstacles to widespread implementation of GPAI in credit scoring lie in the areas of consent and explainability.

[9] Alonso-Robisco, Andrés; José Manuel Carbó (2021). "*Understanding the Performance of Machine Learning Models to Predict Credit Default: A Novel Approach for Supervisory Evaluation*". Documentos Ocasionales 2105, Banco de España.

GPAI achieves optimal performance when processing alternative data sources, which requires explicit, informed consent. It may also rely on real-time data from utilities, payment histories, digital behavior, and employment records—information particularly valuable in assessing *thin credit files*, such as those of young or immigrant applicants with limited financial histories.

Yet, what banks gain in predictive power, they risk losing in transparency. The opaque, *Back-box* nature of many GPAI systems hinders the auditability of decisions, particularly regarding compliance with constitutional anti-discrimination principles that prohibit bias based on gender, age, or origin. Borrowers must be able to understand the rationale behind loan approvals or rejections—a requirement central to both ethical and regulatory frameworks.

Central Banks and Supervisory Authorities

The strong institutional interest in AI extends beyond commercial banking to central banks[10] and supervisory authorities[11], which foresee decisive improvements in prudential and behavioral supervision, payment system monitoring, statistical data analysis, and economic forecasting. GPAI is also expected to enhance emerging functions such as sustainable finance oversight and financial education[12].

GPAI facilitates the verification of soundness, solvency, and institutional conduct to safeguard financial stability. It enables the early detection of latent risks by identifying anomalies and red flags

[10] Cipollone, Piero (2024). "*AI: a central bank's view"*. Keynote Speech at National Conference of Statistics (July 4, 2024). https://www.ecb.europa.eu/press/key/date/2024/html/ecb.sp240704_1~e348c05894.en.html

[11] European Banking Authority – EBA (2023). "*Machine Learning for IRB Models*". EBA/REP/2023/28. https://www.eba.europa.eu/publications-and-media

[12] Lagarde, Christine (2025). "*The Transformative Power of AI*". Welcome address at ECB Conference (April 1, 2025). https://www.ecb.europa.eu/press/key/date/2025/html/ecb.sp250401_1~d6c9d8df11.en.html

in reported data, and it increases the precision of stress-testing simulations used to assess the consequences of adverse economic scenarios[13].

Moreover, GPAI provides superior analytical capabilities for developing early-warning systems for structural imbalances, monitoring payment circuits, and identifying liquidity problems or illicit activities with greater speed. The continuous exploitation of granular, unstructured data allows for the creation of more relevant economic indicators, thereby informing evidence-based public policy[14].

GPAI also strengthens price-discovery mechanisms and reduces barriers to entry in less liquid asset markets, such as corporate debt or emerging markets. Using web scraping and real-time analytics, it can gather and interpret data that reflects both economic and behavioral trends, including information shared through social media.

Improved data quality and utility foster the implementation of automated validation systems to detect errors, outliers, and omissions. This yields substantial savings in model development and training, as well as shorter time-to-market for deployment. ML techniques further contribute by cleaning, inputting, and modeling missing data, thereby improving model robustness and reducing overfitting during training.

In addition, GPAI equips central banks with more effective means of communicating with the public by tailoring their messages linguistically and contextually, expanding accessibility through automatic translation, and simplifying regulatory texts.

---

[13] Balsategui, Iván; Et al. (2024). "*Artificial Intelligence in the banking system: implications and progress from a central bank perspective*". Financial Stability Review - Banco de España, Issue 47, Autumn 2024.
https://repositorio.bde.es/bitstream/123456789/38942/1/1_FSR47_Artificial.pdf

[14] Araujo, Douglas; Et al. (2024). "*Artificial intelligence in central banking: Executive Summary*". BIS Bulletin, 84, Bank for International Settlements. https://www.bis.org/publ/bisbull84.pdf

Finally, by experimenting firsthand with artifiligent applications, central banks can better understand the risks and opportunities of financial innovation, positioning themselves to evaluate its systemic impact and fulfill their own explainability obligations.

Consequently, the banking sector must promptly align with the guidance of central banks and other supervisory authorities, closely monitoring developments such as the *AI Continental Action Plan*[15], in which regulators are expected to play a leading role across several strategic lines defined by the European Commission: GPAI gigafactories; data science laboratories; use-case–based functional architectures; the strengthening of AI-specific talent and skills; and the bolstering of regulatory compliance through simplification and harmonization.

Some Figures

According to data from supervisory authorities, financial institutions, and related stakeholders, the banking sector is undergoing broad and multifaceted adoption of General-Purpose Artificial Intelligence (GPAI).

By late 2024, the *European Banking Authority (EBA) Risk Assessment Questionnaire (RAQ)*[16] reported that most banks in the European Union were employing AI, specifically referencing regression analysis, decision trees, natural language processing, and neural networks.

The report broke down the areas of application and their relative integration rates as follows: mandatory reporting (17.65%); regulatory credit-risk modeling (42.35%); customer service

[15] European Commission (2025). *"AI Continent Action Plan"* https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan

[16] European Banking Authority – EBA (2024). *"Risk Assessment Report / November 2024 – Special Topic: AI"* https://www.eba.europa.eu/publications-and-media/publications/special-topic-artificial-intelligence

(35.88%); credit assessment (54.12%); internal process optimization (60%); anti–money laundering and countering the financing of terrorism (65.88%); fraud detection (69.41%); and customer and transaction profiling and clustering (71.76%).

One year later, the EBA[17] observed that 92% of EU banks had implemented AI systems or methods, while the remaining 8% were conducting pilot projects or evaluating use cases. These initiatives focused primarily on code generation, information extraction and summarization, and the drafting of legal, marketing, or support documentation. Most projects had moved beyond the experimental phase and were being integrated into core ICT infrastructures.

According to the continental supervisor, the most common uses of Artifiligence included: the detection and reporting of fraudulent or suspicious activities; the automation of customer information and guidance tools for self-service digital operations; the deployment of financial education and advisory systems; and the use of digital assistants and call center bots to manage customer service requests.

Across nearly all domains—except regulatory and supervisory reporting—the penetration of AI exceeded 40% compared with traditional tools, reaching approximately 70% in areas such as profiling, internal process optimization, fraud detection, AML, and customer interaction through GPAI or artifiligent agentic systems.

Strategic Approaches

Not all banks have followed the same trajectory in integrating AI, nor have they reached comparable levels of maturity, even regarding standardized GPAI implementation. The sector is simultaneously exploring, researching, and testing diverse strategic models.

---

[17] European Banking Authority – EBA (2025). "*Rising application of AI in EU banking and payments sector*".
https://www.eba.europa.eu/factsheets?text=&document_type=5606&media_topics=All

At the core of this debate lies the choice between developing AI systems in-house, outsourcing development, or integrating third-party models via application programming interfaces, either locally or through cloud infrastructure. Each path carries distinct opportunities and risks, influenced by factors such as scalability, cost, data privacy, cybersecurity, and the availability of qualified personnel.

As banks remain in various stages of experimentation and piloting, this iterative process allows them to assess the strengths and shortcomings of alternative approaches. Once a strategy is chosen, institutions differ in their reliance on proprietary versus open-source systems, or on single versus multiple providers.

Only a small subset of EU financial institutions is currently developing fully proprietary GPAI models, with the high financial and technical hurdles representing the principal constraint. Consequently, European banks appear to favor a multimodal approach, evaluating each business area individually and synthesizing multiple data types and contextual layers. This methodology uplifts security, accelerates processes, supports a holistic (*360-degree*) operational view, and improves system traceability and auditability[18]. GPAI deployment in high-risk areas is generally reserved for situations in which institutions have already achieved a thorough understanding of the underlying models, implementation methods, potential impacts, and mitigation strategies.

Crossroads

Despite the considerable advantages that Artifiligence offers to banking, its integration poses challenges of such magnitude that,

[18] Carbó Valverde, Santiago; Pedro Cuadros Solas, and Francisco Rodríguez Fernández (2023). "*AI and the banking sector: Initial considerations"*. Funcas. *Spanish Economic and Financial Outlook* Vol. 12, No. 4: 33-38.

amid ongoing global uncertainty, it has effectively become a regulated exception within broader technology policies.

This caution stems not only from national prudential concerns, particularly in jurisdictions still sensitive to the repercusions of the *2008 Financial Crisis*[19], but also from a broader continental context. The European Commission's *Apply AI Strategy*[20], which outlines priorities for the next five years, makes no explicit reference to the financial sector. It introduces no flagship initiatives for banking, nor does it directly address workforce training, market-trust mechanisms, or bespoke governance frameworks. Nevertheless, the forthcoming *AI Observatory*, tasked with developing key performance indicators and monitoring AI's evolution, impact, and trends, is expected to maintain an indirect focus on banking-sector GPAI developments.

This omission should not be interpreted solely as anxiety about potential systemic shocks arising from AI deployment in finance. The current global landscape—marked by geopolitical realignment, deregulation, nationalist resurgence, skepticism toward projects such as the *European Banking Union*, resistance to the digital euro, and inconsistent regulation of cryptoassets across jurisdictions—helps explain the sector's slow progress beyond the consultation stage with EU institutions[21].

Ultimately, the adoption of GPAI and artifiligent agentic is governed by a balance of potential and pitfalls. Dependence on external providers increases known IT-related vulnerabilities. GPAI models are currently dominated by a handful of global corporations that

[19] Government of Spain, Ministry of Economy (2024). "*Artificial Intelligence Strategy 2024*". https://portal.mineco.gob.es/es-es/digitalizacionIA/Documents/Estrategia_IA_2024.pdf

[20] European Commission (2025). "*Apply AI Strategy*". Communication from de Commission to the European Parliament and the Council, October 8, 2025. https://digital-strategy.ec.europa.eu/en/policies/apply-ai

[21] European Commission (2024). "*Targeted Consultation on Artificial Intelligence in the Financial Sector*" June 18, 2024. https://finance.ec.europa.eu/regulation-and-supervision/consultations-0/targeted-consultation-artificial-intelligence-financial-sector_en

remain reluctant to submit to European accountability standards or to square with the region's regulatory culture.

Integrating GPAI into banking infrastructure can challenge operational resilience and exacerbate security and privacy risks. While banks are accustomed to managing third-party vendor risk, the complexity and opacity of artifiligent systems amplify these concerns.

Key AI-related threats include model vulnerabilities, such as data poisoning, model tampering, evasion attacks, model reconstruction and extraction, membership inference, backdoor attacks, prompt injection, training data exposure; and malicious use of AI[22]**:** deepfakes, automated malware generation, AI-enabled phishing and social engineering[23], autonomous vulnerability scanning and hacking, AI supply-chain attacks, and denial-of-service operations—among others[24].

Although large-scale AI-driven cyberattacks against the highly fortified cybersecurity systems of major banks remain improbable, the sophistication of such threats fuels public apprehension about data privacy. Citizens increasingly perceive Artifiligence as inherently vulnerable—particularly large language models (LLMs)—to misuse and misinformation[25].

---

[22] European Board for Digital Services (2025). "*First report of the European Board for Digital Services in cooperation with the Commission pursuant to Article 35(2) DSA on the most prominent and recurrent systemic risks as well as mitigation measures*". https://digital-strategy.ec.europa.eu/en/news/press-statement-european-board-digital-services-following-its-16th-meeting

[23] European Commission (2025) "*Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act) Approval of the content of the draft Communication from the Commission - Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)* https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act

[24] The Mitre Corporation (2025). "*Mitre Atlas*". https://atlas.mitre.org/matrices/ATLAS

[25] OWASP (2023). "*OWASP Top 10 for LLM*" Version 1.0. https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-slides-v1_0.pdf

Furthermore, GPAI models rely on vast training datasets that vary in quality, reliability, and privacy safeguards. Most exhibit *hallucinations*—outputs containing inaccurate or misleading information presented as factual. This problem cannot be easily solved without improved data curation, complicating sustainable data governance and heightening litigation risk for financial institutions.

The urgent need for human oversight, both legal and ethical, faces an acute shortage of specialized talent—a gap that can only be bridged through large-scale public and professional literacy in Artificial Intelligence. This, in turn, entails costly training and reskilling programs that many institutions cannot readily afford[26].

These challenges compound traditional banking risks—operational, legal, and reputational—while also introducing a distinct category of *model risk*, which represents both a practical and symbolic frontier in humanity's evolving relationship with intelligent machines.

Crucially, citizens must have access to clear and meaningful explanations of AI's logic and associated risks. The complexity and opacity of artifiligent models render standard disclaimers or interface notices insufficient. Machine-learning systems that operate as a *Black Box* make it challenging to justify decisions with profound human consequences. Accordingly, the *EU Artificial Intelligence Act* imposes stringent requirements regarding governance, transparency, and human oversight.

Training data used in GPAI development can perpetuate discrimination and bias against minority or underrepresented groups, leading to risks of algorithmic financial exclusion. Technical accuracy alone is lacking. It is imperative to design explainable AI

[26] Johnston, Alex; Et al. (2025). "*AI Upskilling: Navigating the urgent need for workforce transformation*". S&P Global. https://www.spglobal.com/market-intelligence/en/news-insights/articles/2025/9/ai-upskilling-navigating-the-urgent-need-for-workforce-transformation-92695030

models that allow banks, regulators, and customers alike to understand and justify algorithmic decisions.

The future of the financial sector will depend, in part, on the ability to reconcile innovation and precision with ethics, transparency, and compliance[27]. At every level, from investor briefings and roadshows to customer interactions, financial institutions must be prepared to explain algorithmic outcomes, whether a high-risk trading position or a denied mortgage application.

[27] Owolabi, Omoshola; Et. al. (2024). "*Ethical Implication of Artificial Intelligence (AI) Adoption in Financial Decision Making*." Computer and Information Science, 17(1), pp. 49-56. https://doi.org/10.5539/cis.v17n1p49

II

BASIS FOR THE EXPLANATION

In the face of rapidly evolving technologies of an inherently disruptive nature, the general principle of Transparency is emerging as a key modulator of *the social right to explanation*. It seeks to counter any form of opacity or abstruseness that may affect fundamental rights, individual freedoms, or the very dynamics of democratic participation.

One salient example is Artificial Intelligence (AI), whether deployed in virtual environments or embodied in autonomous intelligent systems. The imperative for human beings to comprehend the logic underlying any AI-generated output or decision—the structure of its reasoning, its motivation, and its purpose—has crystallized in a series of obligations, techniques, and procedural safeguards commonly known as *eXplainable Artificial Intelligence* (XAI)[28].

Transparency is not a novel concept in the banking sector, which over the past decade has integrated it into various domains of public relevance. This includes the obligation to file annual accounts; compliance requirements in advertising and marketing; the provision of pre-contractual documentation for mortgage applications; the duty of double transparency in contractual matters to prevent unfair terms; and supervisory reporting obligations focused specifically on transparency. The same ethos of openness has increasingly extended

[28] Cotino Hueso, Lorenzo; Jorge Castellanos Claramunt (editors) (2022). "*Transparency and Explainability of Artificial Intelligence*". Tirant Lo Blanch, Valencia, 2022. https://www.uv.es/cotino/publicaciones/libroabiertotp22.pdf

to relationships with outsourcers[29] (*open book*) and with customers (*open banking*)[30].

The purpose of XAI is to translate algorithmic justifications into explanations comprehensible to non-specialist audiences. Interpretability—the internal coherence of the model—is of negligible value if intelligibility—the user's understanding—remains poor. Crafting explanations, including their structure, format, and linguistic framing, must therefore adapt to the recipient's cognitive and cultural context.

This challenge is magnified by the advent of Large Language Models (LLMs). While LLMs can assist in generating explanations, their complexity obstructs the traceability of reasoning. A subtle issue, such as linguistic bias in English-dominated training data, can lead to misrepresentations or reduced accessibility in other languages.

At its core, this is a question of expression and comprehension, an intricate problem of translatability between languages and even epistemic paradigms—akin to those raised by hieroglyphics, allegory, Kabbalah, or cryptography. In essence, it is a problem rooted in semantics, and ultimately in Epistemology[31].

This raises a fundamental tension: whether to impose limits on research to ensure intelligibility, or to accept that certain cognitive depths of Artifiligence will remain opaque. Addressing this issue requires engagement with linguistic and philosophical conventions about meaning and interpretation—an uncomfortable but unavoidable endeavor.

---

[29] Crown Commercial Service [United Kingdom] (2016). "*Open Book Contract Management – OBCM Guidance*" https://assets.publishing.service.gov.uk/media/67b480483e77ca8b737d37be/Open_book_contract_management_guidance.pdf

[30] Duncan, Ellie (2024). *"Open Banking and Financial Inclusion*". Kogan Page Ltd.

[31] A vivid memory from my childhood—In a small room attached to the Securities Department, at the headquarters of the *Savings Bank of La Coruña and Lugo*, my father silently deciphers, engrossed, the punched tapes with the stock market quotations that have arrived by telex, after five in the afternoon, once the *Madrid Stock Exchange* has closed.

In relation to XAI terminology, the scientific community continues to debate definitions of *interpretability* and *explainability*, often emphasizing their distinctions. In corporate practice, even among institutions managing high-risk AI systems (HRAIS), these terms are rarely used beyond system-engineering teams, and even there, inconsistently. While linguistic consensus may eventually emerge, from the client's standpoint, we adopt the following conceptual distinctions in this study.

Transparency, in its strict sense, refers to the obligation to inform clients that a decision is automated and, where applicable, that they are interacting not with a human agent but with an intelligent system. *Explainability* refers to the methods used to generate human-understandable justifications for predictions or classifications made by complex or opaque (*Black Box*) models.

Interpretability, for engineers, denotes a model's intrinsic capacity to be understood immediately without auxiliary tools; for clients, however, it concerns their subjective grasp of the explanation. In many languages, *interpretability* even carries a pejorative connotation, implying ambiguity that may advantage the contractual party with superior interpretive clout.

Whereas *transparency* may be satisfied through formal acknowledgment of information received, *interpretability* and *explainability* reflect the heterogeneous nature of human understanding. The core risk for banks, therefore, lies not only in *the adequacy of the explanation furnished* but in *the likelihood of misinterpretation by the recipient.*

Moreover, *explainability* refers to the property, skill, or capacity of a system to be explainable, whereas *explanation* denotes the explanatory content provided to clarify how or why a system arrived at a given outcome.

The question is far from idle[32]. Even before the popularization of AI-Gen, prominent scholars and industry leaders argued that explainability should remain an ancillary requirement—valuable, but not essential to the transformative paradigm shift brought by Artificial Intelligence, and acceptable only insofar as it does not impede technological development[33].

Today, however, we face an inescapable trade-off between predictive accuracy and interpretability. It is an irrefutable fact that the most advanced and accurate artifiligent models are becoming increasingly opaque. Simplifying a model to make it more comprehensible often involves a decline in performance, which, in critical domains—such as medical diagnosis, autonomous driving, or fraud detection—may produce outcomes detrimental to the user, making certain forms of XAI potentially counterproductive[34].

Another argument frequently raised concerns the ease with which an illusion of explanation may arise[35]. The mere perception of understanding can masquerade as genuine comprehension, fostering unwarranted confidence. Explanatory narratives may conceal fallacies embedded in the model's design[36]. Many contemporary XAI

---

[32] European Data Protection Supervisor - EDPS (2023). "*TechDispatch. Explainable Artificial Intelligence*" https://www.edps.europa.eu/system/files/2023-11/23-11-16_techdispatch_xai_en.pdf

[33] Lipton, Zachary C. (2016). "*The Mythos of Model Interpretability*". ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY https://arxiv.org/pdf/1606.03490.

[34] Rudin, C. (2019). "*Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*". *Nat Mach Intell* **1**, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

[35] Bhatt, Umang; et al. (2020). "*Explainable Machine Learning in Deployment*" Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency: 648-657. https://arxiv.org/pdf/1909.06342

[36] Selbst, Andrew D.; Et al. (2019). "*Fairness and Abstraction in Sociotechnical Systems"* FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), ACM New York. https://doi.org/10.1145/3287560.3287598

methods operate *post hoc*, generating simulations or interpretive reconstructions that do not reveal internal mechanics but instead provide reflections or narratives that appear plausible while failing to represent the real decision-making process of the system[37].

This trend is well documented in psychology, particularly in an era marked by misinformation and post-truth currents: individuals are prone to illusions of causality, simplicity biases, and susceptibility to persuasive explanations that become dogmatic despite being inaccurate. As a result, explanations may distort rather than illuminate the underlying logic of the model.

The challenge is compounded by the absence of standardization and the impracticability of imposing a universal explanatory pattern, as expectations vary across individuals and contexts. Unlike model accuracy, which can be evaluated with objective metrics, the quality of an explanation is inherently subjective and context-dependent[38].

Detailed explainability frameworks also introduce legal and strategic risks. If an explanation discloses a spurious correlation, a latent bias, or a technical flaw, financial institutions may incur heavier liability than if the decision-making process had remained inscrutable. This dynamic may discourage transparency and incentivize organizations to prioritize token adherence instead of genuine operational change. Debates over what exactly constitutes the *explanandum* add further complexity to this dilemma.

Nonetheless, XAI is indispensable for fostering trustworthy Artifiligence. Beyond its relevance for debugging models, explainability plays a critical role in mitigating biases, enabling

---

[37] Ghassemi M; Luke Oakden-Rayner and Andrew L. Beam (2021). "*The false hope of current approaches to explainable artificial intelligence in health care*". Lancet Digit Health. 2021 Nov;3(11):e745-e750. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00208-9/fulltext.

[38] Wachter, Sandra; Brent Mittelstadt and Chris Russell (2017). "*Counterfactual Explanations Without Opening the Black Box*". Harvard Journal of Law & Technology, https://arxiv.org/pdf/1711.00399

traceability and auditability, and fostering public trust, reducing the sense of unease that often attends technological innovation.

Legal Framework

Although a laudable process of global normalization—and even standardization—is underway, nothing can be taken for granted amid the current geopolitical volatility, one of whose hallmarks is the rise of a *deregulation guerrilla* in certain jurisdictions.

What can be evidenced, however, is a clear trend. Over the past five years, a diversity of soft-law approaches has emerged—ranging from multilateral frameworks developed by major international organizations (UNESCO[39]; OECD[40]) to regional or national initiatives rooted in local regulatory traditions. These approaches have converged toward a prescriptive, auditable model grounded in shared principles and practical implementation guidelines.

In the United States, the National Institute of Standards and Technology (NIST)—a public agency within the Department of Commerce—first published the *NIST AI Risk Management Framework (AI RMF)*[41], which recommends incorporating explainability throughout the AI product lifecycle and providing documentation and confidence metrics. This was followed by the *NIST AI RMF GAIP*, specifically addressing Generative AI[42], and

[39] UNESCO (2021). "*Recommendations On the Ethics of Artificial Intelligence*". https://unesdoc.unesco.org/ark:/48223/pf0000381137

[40] Lorenz, Philippe; Karine Perset and Jamie Berryhill (2023). "*Initial policy considerations for generative artificial intelligence*", *OECD Artificial Intelligence Papers*, No. 1, OECD Publishing, Paris, https://doi.org/10.1787/fae2d1e6-en.

[41] National Institute of Standards and Technology-NIST (2023) [USA]. "*Artificial Intelligence Risk Management Framework*" https://www.nist.gov/itl/ai-risk-management-framework

[42] National Institute of Standards and Technology - NIST (2024) [USA]. "*Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile".* https://doi.org/10.6028/NIST.AI.600-1

subsequently by President Biden's *Executive Order on AI*[43], partially revoked by President Trump's *Executive Orders* of January 23, 2025[44], and December 12, 2025[45], consistent with his imperialist doctrine and the anarcho-capitalist ideology of the libertarian oligarchy.

In China, the programmatic *Beijing AI Principles*[46] are being operationalized through targeted regulations (e.g., deepfake governance). Similar developments are unfolding in jurisdictions such as Japan[47] and Brazil[48].

The United Kingdom deserves particular attention. Despite pressure from political parties or sectoral lobbies, it continues to favor a principle-based regulatory model—structured around *Safety, Security and Robustness; Transparency and Explainability; Fairness; Accountability and Governance; and Contestability and Redress*[49]. This approach is explicitly pro-innovation and sector-specific[50], with responsibility for banking-related AI residing in the

---

43 Biden, Joe. *Executive Order 14110, titled Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (2023). https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

44 Trump. Donald. J. *Executive Order 14179 of January 23, 2025 (Removing Barriers to American Leadership in Artificial Intelligence)*

45 Trump, Donald J. *Executive Order 14365 of December 11, 2025 (Ensuring a National Policy Framework for Artificial Intelligence)*

46 Beijing Academy of Artificial Intelligence (BAAI) [China]; Et al. (2019). "*Beijing AI Principles*" https://ai-ethics-and-governance.institute/beijing-artificial-intelligence-principles/

47 Government of Japan (Cabinet Office) (2018) [Japan] "*Social Principles of Human-Centric AI*" (2018). https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf

48*Projeto de Lei No. 2338/2023, which dispõe sobre o uso da Inteligência Artifici* (November 2025). https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

49 Leslie, David [The Alan Turing Institute]; Information Commissioner's Office-ICO (2020). "*Explaining Decisions Made With AI"*
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4033308

50 UK Government, Department for Science, Innovation & Technology (2024). "*A Pro-innovation approach to AI Regulation (Government response to consultation – White Paper*". https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

Financial Conduct Authority (FCA), which has already begun to articulate supervisory expectations[51].

In contrast, the European Union has adopted what is commonly referred to as the *AI Brussels Standard*, reflecting the EU's regulatory tradition: binding, risk-based legislation, currently embodied in two instruments of direct application across the Union—the *General Data Protection Regulation* (GDPR)[52] and the *Artificial Intelligence Act* (AI Act)[53].

The complex interaction between these two regulatory spheres will not be examined here. Until domestic supreme courts, or ultimately the Court of Justice of the European Union (CJEU), consolidate authoritative case law, interpretation will continue to rely on the guidance of supervisory authorities—drawing on precedents established in other areas (e.g., credit files[54]) while many AI-specific supervisory bodies are still emerging—as well as on scholarly research[55].

In broad terms, the GDPR enshrines the right to an explanation, while the AI Act delineates the technical and organizational requirements to guarantee that such explanations are rooted in a

---

[51] Financial Conduct Authority – FCA [United Kingdom] (2025). "*FCA AI Update*" https://www.fca.org.uk/publication/corporate/ai-update.pdf

[52] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)* https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[53] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828.* https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[54] Spanish Data Protection Agency - AEPD (2019). "*Consultation 028891/2019*" https://www.aepd.es/documento/2019-0081.pdf

[55] Gavara Feijóo, Pablo; María Piedrafita Abión [APDCAT] (2024). "*GDPR vs. RIA. Analysis of a partial insertion*". https://apdcat.gencat.cat/web/.content/03-documentacio/estudis-recerca/RGPDvsRIA_es.pdf

secure, robust, and fair system. Compliance, therefore, is not secured solely by generating a *post hoc* explanation; the technical foundations of the high-risk model—data quality, robustness, and traceability—must meet regulatory standards. A failure in robustness or an inadequately mitigated bias may render even the most meticulously crafted explanation misleading, invoking regulatory penalties.

A/ GDPR

Article 22 GDPR grants the right of the data subject to remain free from a decision based solely on automated processing, including profiling, where such a decision produces legal effects concerning him or her or otherwise significantly impacts the individual. Although the right is limited to fully automated decisions, in exceptional cases, the GDPR mandates controllers to institute appropriate safeguards, including the right to obtain human intervention, to express one's point of view, and to contest the decision.

Articles 13–15 GDPR demand a functional interpretation, compelling controllers to provide the data subject with substantive insight into the model's rationale**.** While EU rulebook does not recognize a freestanding *fundamental right to an explanation*, the wider array of rights that define the right to a fair trial—constitutionally protected throughout Europe—includes a firmly-rooted mandate for reasoned decisions. This governs both public authorities and any natural or legal person whose actions may infringe other fundamental rights, such as equality and non-discrimination on the grounds of birth, race, sex, religion, opinion, or any other personal or social circumstance.

To uphold the citizen's right to challenge an automated decision requires that the technical explanation be translatable into a legally defensible, auditable argument. This legal imperative elevates the strategic importance of transparency or explanations exhibiting high explanatory fidelity, so that the logic presented to affected individuals

is coherent and amenable to external scrutiny. Explainability thus becomes a pivotal cornerstone element of legal auditing in AI.

The caselaw of European national data protection authorities, addressing automated decision-making (ADM) under Article 22 GDPR, centers on protecting individuals from decisions taken without meaningful human involvement[56]. A purely automated decision is one executed exclusively by technological means, lacking consequential human input. The prohibition in Article 22 GDPR applies only if the decision stems exclusively from automated processing. Profiling—defined as processing aimed at evaluating personal aspects such as work performance, economic situation, or health—often overlaps with automated processing, though the concepts are not identical.

The prohibition targets only decisions that invoke legal consequences (e.g., refusal of credit or employment) or that significantly affect the individual (*le afecte de manera significativa*; *l'affectant de manière significantly similaire; in ähnlicher Weise erheblich beeinträchtigt*). The term *significantly* embodies an inherently vague legal concept (*Unbestimmter Rechtsbegriff*) that courts and supervisory authorities must clarify. It broadly encompasses non-trivial impacts on access to services or opportunities, impediments to freedom of movement or decision-making, and forms of exclusion or discrimination, even absent illegality.

Automated processing is waived from prohibition and deemed lawful, only when one of three exceptions obtains: a) the processing is necessary for entering into or performing a contract; b) the processing is authorized by EU or Member State law; or c) the data subject has provided explicit, informed consent. This final exception

[56] Spanish Data Protection Agency – AEPD (2024). "*Right not to be subject to automated individual decisions*". https://www.aepd.es/derechos-y-deberes/conoce-tus-derechos/derecho-no-ser-objeto-de-decisiones-individuales

warrants particular scrutiny, given widespread semi-automatic acceptance practices online and frequent information asymmetries in telemarketing, contractual terms, and consumer interfaces.

Even where an exception prevails, controllers must uphold: (i) the right to human intervention; (ii) the right to express one's view; and (iii) the right to contest the decision. To preclude an ADM from being categorized as "solely automated", human intervention must be substantial, demanding[57]: a) that the human reviewer possess the competence and legal or hierarchical authority to alter the outcome, along with adequate training to assess the decision, the underlying data, and the system's capabilities and limitations; and b) that the review embraces a genuine assessment of all relevant data, including additional information provided by the data subject. Superficial reviews, rubber-stamping of outputs, or interventions thwarted by resource deficits do not fulfill this requirement**.**

Automated processing intensifies the controller's duties of transparency and proactive accountability. Articles 13–15 GDPR explicitly mandate notifying data subjects about the existence of automated processing, including profiling, the logic involved, and its intended consequences. Automated processing that involves systematic and extensive evaluation of personal aspects, which in turn produces decisions with legal effects, necessitates a *Data Protection Impact Assessment* (Article 35 GDPR). Processing involving special categories of data (e.g., health or racial origin) is proscribed unless explicit consent is obtained or specific exceptions apply, accompanied by appropriate safeguards (Article 9 GDPR).

[57] Spanish Data Protection Agency – AEPD (2024). "*Evaluation of human intervention in automated decisions*". https://www.aepd.es/prensa-y-comunicacion/blog/evaluacion-de-la-intervencion-humana-en-las-decisiones-automatizadas

Credit scoring falls squarely within these GDPR provisions[58]. In its judgment of 7 December 2023 (*Schufa*)[59], the CJEU ruled that the calculation and transmission of a credit score by a commercial agency to a bank's risk manager transcends a mere preparatory act to constitute an automated decision with legal effects. Consequently, the data subject's rights under Article 22 apply to the scoring agency itself[60].

Non-adherence triggers the specific GDPR sanctions regime, supplemented by domestic administrative sanctions law—substantive and procedural—and is governed by the overarching principles of legality, non-retroactivity, culpability, proportionality, presumption of innocence, *non bis in idem*, prescription, and procedural guarantees such as the right to be heard, access to evidence, legal assistance where appropriate, and a decision within a reasonable time.

B/ AI Act

The AI Act augments the GDPR by delineating the technical and organizational protocols required to ensure that explanations are robust, secure, transparent, traceable, and non-discriminatory. It implements a risk-based approach that segregates: a) *unacceptable risk* systems—proscribed for jeopardizing security, fundamental rights, or economic stability (e.g., social scoring or policing

[58] Campos Rivera, Gonzalo. "*Credit Scoring as the processing of personal data in light of the GDPR* (2024). UNED Law Review, no. 33/2024. https://revistas.uned.es/index.php/RDUNED/article/view/41926

[59] Court of Justice of the European Union (2023). *Judgment of 7 December 2023 (Shufa Case)* https://curia.europa.eu/juris/document/document.jsf;jsessionid=32730EC98571CACB1615E69882702DEE?text=&docid=280426&pageIndex=0&doclang=ES&mode=req&dir=&occ=first&part=1&cid=489949

[60] Guerrero Ovejas, Marta (2024). "Automated scoring in *credit granting"*. Working Paper 5/2024 – Jean Monnet Chair. https://diposit.ub.edu/dspace/bitstream/2445/214987/1/WP%20Marta%20Guerrero%20Ovejas_2024_5-1.pdf

predictions based solely on profiling); b) *acceptable risk* systems, divided into (i) *limited-risk* and (ii) *high-risk* AI systems (HRAIS), deployed in sensitive domains such as critical infrastructure, employment, education, finance, law enforcement, and Justice; and c) *minimal risk or no risk* systems.

Transparency and explainability obligations stem directly from this taxonomy. For *limited-risk* systems, these mandates are slight: users must merely be apprised that they are engaging with an artifiligent system (e.g., chatbots, text generators).

For *high-risk* systems, mandates intensify considerably: a) transparency and explainability, demanding that models achieve inherent transparency or employ appropriate rendering methods; b) strict data governance, upholding data veracity, relevance, and crucially, the mitigation of algorithmic bias; c) robustness and accuracy, compelling systems to demonstrate resilience and curb cascading failures; and d) effective human oversight during deployment and monitoring to prevent harmful outcomes.

In broad terms, the AI Act establishes global explainability duties on the *provider/developer* directed at the *deployer*, while the *deployer* must ensure local explainability to the *end user*, who must be notified upon exposure to HRAIS (Art. 14 AI Act) and whose right to human supervision must be secured (Art. 27 AI Act).

Article 13 AI Act mandates granular transparency on providers of HRAIS to foster intelligibility for deployers. Providers must engineer systems transparently to enable deployers to comprehend system operation and resultant outputs. They must furnish exhaustive *Instructions for Use*, detailing: intended purpose; limitations; output interpretation methods; human-monitoring procedures; and protocols for the proper collection, storage, and interpretation of system logs.

However, ambiguities are not fully resolved by the AI Act: the identity and role of those participating in the explainability chain (explainers and recipients); the nature of the explicandum; and the techniques appropriate to generate acceptable explanations in

contexts of inherent model opacity. Subsequent supplementary frameworks—such as the *GPAIMs*[61] and the *Code of Practice on Transparent AI Systems*[62]—introduce a further actor, the *downstream modifier*, who customizes or refines models and may, under specific conditions, assume the role of *provider/developer* with concomitant liabilities.

The field is seeing the emergence of a corpus of case law and scholarship, which will be capped by the sanctions regime. A major challenge will be the apportionment and synergy of supervisory mandates, given the proliferation of financial supervisory authorities at both the EU and national levels. These notably comprise data protection authorities (EDPB), AI supervisory bodies (EU AI Office; ENISA), or banking supervision (EBA; ECB; SSM; ESRB), as well as others rapidly multiplying at the domestic level[63].

[61] European Commission (2025). GPAIMs "*Communication to the Commission Approval of the content of the draft Communication from the Commission – Guidelines on the scope of the obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act)* " https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers

[62] In drafting phase so far (January 2026) https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content

[63]European Commission (2025). "*Digital Omnibus (Draft) / Rationalising the governance by granting the AI Office more oversight / November 2025 Draft*". *https://digital-strategy.ec.europa.eu/en/policies/digital-rulebook*

III

STRUCTURE OF THE EXPLANATION

Who *should explain*

Liability for explainable artificial intelligence (XAI) within general-purpose AI (GPAI) systems, from the perspective of European Union regulation, is governed by the specific role each actor occupies through the value nexus of creation, implementation, and use, as well as on the risk level of the AI system concerned. This interplays with contractual liability arising from service agreements between the various participants in the value chain and, *de lege ferenda*, tort liability for damages[64].

The *provider* or *developer*—namely, the legal person that conceives, constructs, and delivers the base model—must affirm the legality of the data used and furnish accurate information regarding the model's capabilities, thereby securing transparency and copyright compliance. Providers must draw up and maintain technical documentation, submit detailed reports on training data, and collaborate with European authorities. These mandates are significantly amplified when the GPAI system poses a systemic risk—defined as attaining a computational capacity threshold of 10^{25} FLOPS—compelling comprehensive risk and mitigation assessments, reporting of serious incidents, and the deployment of stringent cybersecurity measures to shield end users from large-scale risks associated with the most powerful models. Additionally,

[64] European Commission (2022). "*Proposal for a Directive of the European Parliament and the Council on adapting non-contractual civil liability rules to artificial intelligence*" *(2022)* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022PC0496

providers/developers are obligated to remit their XAI packages to the EU AI Office.

Within this liability regime, *downstream modifiers* must be pinpointed whenever alterations are classified as *substantial* under Article 3(23) of the AI Act—namely, changes not explicitly envisioned or planned in the original conformity assessment. Examples encompass: a change of purpose that transforms a limited-risk system into a high-risk AI system (HRAIS); the tuning of hyperparameters in ways that exacerbate bias; material shifts to the system's architecture; changes that compromise robustness or cybersecurity; or modifications to the data flow. Conversely, the following do not constitute substantial modifications: updates to software libraries; security patches; bug fixes; changes to user interfaces; or adjustments to alert thresholds within the predefined technical framework.

The *deployer* is the legal entity (public authority, corporation, or company) that integrates a GPAI model or an HRAIS into a final product or service that engages end users; for example, a bank implementing GPAI into a credit assessment system. The AI Act distinguishes: (i) deployment of an HRAIS: the *deployer* assumes primary accountability toward the user, including mandates to ensure human oversight; verify input data integrity and mitigate bias; maintain system logs for auditability; and guarantee use of the system in conformity with the *Instructions for Use*. The *deployer* is directly liable for ensuring that the final application or interface—i.e., the system rendering the consequential decision—is fair, accurate, and properly supervised; (ii) deployment of *limited-risk* systems: not all AI-supported banking activities are deemed high risk; in these cases, *deployers* must adhere to basic transparency mandates, such as informing users when interacting with an AI system (e.g., a chatbot).

Other actors may also incur XAI-related liability. Any substantial modification subsumes them under the *provider/developer* liability regime; moreover, specific duties may be generated depending on

how their actions bear upon explainability. This pertains to *importers* and *distributors*, who must ascertain the provider's compliance with EU regulations and guarantee that storage conditions do not compromise those obligations. *External parties* that tailor or transform systems may likewise assume liability.

On this legal foundation of XAI responsibility is erected the architecture of contractual liability, which holds sway insofar as it does not impinge upon *jus cogens*—mandatory legal norms whose content cannot be excluded or modified, rendering any contrary stipulation null and void.

In the evolving landscape reshaped by Artificial Intelligence, traditional contracts governing New Technologies, used until now to regulate private relationships born of the Digital Revolution, retain only residual relevance. These comprise leasing agreements; software acquisition and distribution contracts; user licensing agreements; IT outsourcing; systems integration; backup arrangements; database transfer contracts; Electronic Data Interchange; Peering Agreements; Network Access Agreements; and Internet Reach and Market Agreements.

A distinct category of contracts is now taking shape[65], featuring legal structures that have been met with resistance by practitioners and scholars, as is the case with *algorithmic contracts*[66], although certain types are already charting a course toward global standardization.

The *Artificial Intelligence as a Service* (AIaaS or Model-as-a-Service) contract, widely used for commercial Large Language

[65] Ebers, Martin; Et al. compilators (2022). "*Contracting and Contract Law in the Age of Artificial Intelligence*". Bloomsbury.

[66] Scholz, Lauren Henry (2017). "*Algorithmic Contracts*", Stanford Technology Law Review 128.https://law.stanford.edu/wp-content/uploads/2018/03/3_SCHOLZ-FINAL_Formatted_Mar18.pdf

Models (LLMs), affords clients access to a model's capabilities through an API or web interface. Because ownership of the base model rests unequivocally with the *provider/developer*, contractual clauses center primarily on disclaimers of liability and the scope of authorized use of the output generated by the system.

The *Foundational Model License Agreement* (*On-Premise*) entails the *provider/developer* furnishing a replica of the essential components of the model for installation within the client's infrastructure, who then assumes the role of deployer-licensee. Functionally, this is a software and intellectual property licensing contract. This model is typically adopted by organizations requiring full control over their data and system performance. Responsibility is shared but may migrate partially or fully to the licensee, depending on whether they introduce minor adjustments or substantial modifications.

The *Fine-Tuning Contract*, sometimes accompanied by a retraining agreement, governs scenarios where *the provider/developer* or a third party customizes a foundational model (e.g., an LLM) using the client's proprietary data to calibrate the system to a specific operational domain, such as banking. Here, it is imperative to clearly define ownership of the training inputs (the dataset owner), the confidentiality framework enveloping such data, and ownership of the resulting model.

In HRAIS procurement agreements, which are particularly relevant in the public sector or regulated industries such as finance, the contract constitutes either a supply or service provision agreement supplemented by the AI Act's regulatory requirements (risk management, documentation, traceability, accuracy, etc.). At a minimum, such contracts must entitle the client to audit the system at any time.

In the realm of intellectual property, AI exclusion clauses (*opt-out*) have burgeoned: creators and rights holders contractually proscribe or constrain the use of their digital works for training GPAI

models (text and data mining), instituting a demand for royalties when such material is used.

Given the current legal uncertainty, contracts may also articulate ownership of outputs, explicitly determining whether outputs generated by an LLM appertain to the *provider/developer*, the *downstream modifier*, the *deployer*, the user (*prompt engineer*), or the public domain. Related issues include whether providers may leverage user interactions for model improvement or retraining, and how confidential data must be handled upon service termination.

Contracts increasingly incorporate warranties and indemnities confirming that: (i) the training data was lawfully secured; (ii) the *provider/developer* holds the necessary rights or licenses to use such data; (iii) the AI system has been designed and tested to yield accurate results and minimize material bias to the extent reasonably possible; and (iv) the *provider/developer* will undertake continuous testing to monitor accuracy, safety, and system suitability.

A crucial yet contentious area—still largely evaded by major law firms representing U.S. technology giants—is the liability regime for incorrect, misleading, false, or fabricated information produced by AI systems (*Limitation of Liability for Errors and Hallucinated Content*). *Provider/developer* typically endeavors to constrain or foreclose liability, whereas counterparties aim to impose minimum accuracy guarantees or require compensation for losses resulting from erroneous outputs, including exposure to legal peril.

Until a clear and strict legal framework for non-contractual liability in AI is codified, courts will be compelled to adjudicate disputes by relying on general principles of contract law, the terms of professional or cyber liability insurance covering AI-related risks and

third-party claims, and the existing framework of strict liability for damages caused by defective or evolving products[67].

It is also worth noting that we are on the cusp of encountering fully developed autonomous robotic systems, integrating advanced robotics and deep AI, operating in physical form, whether anthropomorphic or otherwise. European institutions are already investigating how to regulate liability for these sophisticated devices[68]. This process requires initially establishing the legal foundations governing their creation, implementation, and operation within the European Union, including their legal status, a taxonomy of relevant facts and harms, and appropriate methods for redress and resolution of damages[69].

Explaining What *to* Whom*: The Five Beacons*

The primary party safeguarded by European regulations on explainable artificial intelligence (XAI) is the end user—in this case, the banking customer—to whom the reasons underlying a decision (*Reasons Code*) rendered by a non-human entity must be substantiated, for which the bank bears the direct burden of proof.

However, fulfilling this requirement in a matter of such technical and legal complexity necessitates structuring user protection across a set of intermediate actors, each serving as an anticipatory safeguard—elucidations designed to preclude individuals from succumbing to informational vulnerability in the face of the machine.

---

[67] *Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC.* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202402853

[68] *European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).* (2017). https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html

[69] Verbovaya, Olga; Et al. (2024). "*Responsibility for damages done by robots: The European Union experience*". *Sci Herald Uzhhorod Univ Ser Phys.* 2024;(55):1792-1801. DOI: 10.54919/physics/55.2024.179ge2

Five beacons are distinguished, each associated with a different level or form of explainability. What must be explained (*explanandum*) varies in scope and depth depending on the intended recipient, who embodies a distinct dimension of the human interest in being informed about the logic guiding decisions made by computational artifacts.

Rather than treating explainability as a single explanation delivered at the end of an automated process, this model establishes a sequence of prior explanations, each addressed to actors capable of understanding and validating increasingly complex aspects of the system's behavior. This structure ensures that the explanation ultimately provided to the customer is supported by earlier layers of technical, organizational, and supervisory accountability.

A/ XAI-M2M

In M2M XAI (*Machine-to-Machine Explainability*)—whose practical applications are rapidly proliferating within the Internet of Things (IoT)[70]—systemic explainability pertains to the conditions that enable numerous distributed applications or autonomous agents to interoperate seamlessly and consistently without loss of efficiency, coherence, or robustness, and without security breaches within complex networks and critical infrastructures, including those governed by Zero Trust Architecture (ZTA), the cybersecurity paradigm founded on perpetual distrust of every internal and external system component.

This gives rise to a set of interoperability requirements: the explicability capacity of the model; specification of the XAI techniques employed; the adoption of standardized explanation

[70] Favour, Akintan; Et al. (2025). "*Explainable AI Models for Trust Evaluation in M2M Communication Security Systems*" https://www.researchgate.net/publication/393408348_Explainable_AI_Models_for_Trust_Evaluation_in_M2M_Communication_Security_Systems

formats; data-flow traceability; detailed and persistent logging; data transparency; fungibility of explanations; modularity of inference components; the principle that the explainability layer must remain independent of underlying communication architectures (*protocol agnosticism*); computational efficiency to avert impeding critical M2M operations; acceptable latency; and stringent security and integrity safeguards to thwart the manipulation of explanations. Compliance with the regulatory standards applicable in the relevant jurisdiction is, of course, mandatory.

In contrast to the instantaneous acceptance-or-rejection logic characteristic of machine interconnection, M2H XAI (Machine-to-Human Explainability) opens a broad field of variability, shaped by the inherent diversity and interpretive ambiguity of the human mind. No explanation—however objective it may appear—will be interpreted identically by two humans. Within this domain, four typical XAI recipients emerge, each associated with distinct levels of explanatory demand: *XAI-Engineer*; *XAI-Supervisor*; *XAI-Corporation*; and *XAI-Client*.

B/ *XAI-Engineer*

The Engineer—a category encompassing professionals employed by deployers or modifiers, external system auditors, independent assessors, or court-appointed experts—requires detailed technical explanations that empower them to debug, evaluate technical resilience, assess security postures, and optimize model performance by understanding how a specific output was generated.

This form of explainability is delivered primarily through technical documentation that enables replication, review, and, where appropriate, further development of the system. Following their analysis, engineers must furnish a concise report including a *certificate of conformity* or, where appropriate, a grounded exposition

of non-compliance and the scope of any objections, reservations, or caveats.

Regarding system design and project purpose: (i) *Machine Learning Canvas*, which may be visual or textual, must include: the business objective optimized by the AI system; the success metric, with numerical KPIs; the type of model (classification, regression, clustering, etc.); the deployment and prediction context (e.g., real-time API, daily batch); and the regulatory restrictions governing its XAI; (ii) *Data Card*, describing the dataset used to train and evaluate the model: data origin, sampling and date of collection; a list of all features with their definitions, data types, and value ranges; basic statistics, including target distribution and missing-data analysis; and identified limitations and biases (e.g., historical bias in legacy data) and their potential impact.

Regarding technical documentation: (iii) *Model Card*, detailing system construction: algorithm type and rationale; preprocessing steps (e.g., normalization, one-hot encoding, outlier treatment); feature engineering; hyperparameters and their values; evaluation results (accuracy, recall, precision, F1-score, AUC) in both test and validation sets; and an explanation identifying the most important variables (feature importance); (iv) *Code and Environment Repository*, including version-controlled source code for preprocessing, training, evaluation, and deployment; and a dependencies file specifying library and framework versions.

Regarding production and maintenance: (v) *Deployment Documentation*, specifying how the model is accessed (API, interface), latency and performance expectations, and hardware/software requirements; (vi) *Monitoring Documentation*, explaining how model drift and decay are detected, thresholds triggering alerts, and procedures and schedules for updates and retraining.

C/ *XAI-Supervisor*

Supervisors necessitate transparency centered upon traceability. Full access to the system's life-cycle documentation is essential for verifying compliance with legal standards and the effectiveness of bias-mitigation processes.

Under the European Central Bank's (ECB) Single Supervisory Mechanism (SSM), national banking supervisors audit AI systems from a prudential and operational-risk perspective. Until sector-specific regulation crystallizes, these audits must rely on established data and model governance frameworks, suitably adapted to the specificities of Artificial Intelligence[71].

In terms of AI Governance and Strategy, which sets the high-level framework for oversight: (i) *Governance Policy*, describing roles and responsibilities (e.g., Chief AI Officer, AI Ethics Committee); the model inventory with criticality and function classifications (e.g., pricing, credit scoring, AML); and the Risk Appetite Framework, with explicit limits for accuracy, bias, and stability: (ii) *Internal Validations*, including the Model Validation Report (assessing suitability, performance, and stability) and the Internal Audit Report (evaluating the effectiveness of development, deployment, and monitoring processes).

Regarding development and life-cycle documentation: (iii) *Model Design Document*, outlining the business purpose, prediction task, and value proposition; the rationale for choosing a specific AI algorithm (e.g., Deep Learning vs. Random Forest); and an assessment of the model's explainability, especially for black-box

[71] European Banking Authority – EBA (2025) "*AI Act: implications for the EU banking and payments sector*" https://www.eba.europa.eu/sites/default/files/2025-11/d8b999ce-a1d9-4964-9606-971bbc2aaf89/AI%20Act%20implications%20for%20the%20EU%20banking%20sector.pdf

systems. (iv) *Data Management Report*, detailing internal and external data sources, quality, cleaning and transformation processes, bias and fairness analysis, and data-retention and traceability policies (data lineage).

Regarding monitoring and deployment documentation: (v) *Monitoring Report*, comparing production metrics with testing-phase benchmarks; identifying data drift or concept drift; and listing alerts or automated corrective actions; (vi) *Retraining and Versioning Specifications*, describing retraining procedures, version history, and rationales for updates; (vii) *Evidence of Resilience and Cybersecurity*, documenting adversarial testing and security measures for the hosting infrastructure.

Supervisors further demand documentation on non-financial risks, which are particularly acute in the AI domain: (viii) *Impact on Fundamental Rights*, mandating the absence of discriminatory impact on protected groups and justifying the level of human intervention; (ix) *Legal and Compliance Risk Report*, confirming adjustment with applicable regulations, particularly when the system qualifies as HRAIS; (x) *Compilation of Reasons Code* and representative explanation files; (xi) *Cross-references to the other XAI Beacons*, ensuring alignment across the entire explainability structure.

Following this review, the supervisory authority prepares a report containing findings and conclusions, which is generally not made public. In cases of disagreement, disciplinary proceedings may be commenced under the relevant substantive and procedural rules.

D/ *XAI-Corporation*

Bank personnel, group encompassing employees of outsourcing firms, also require XAI. Corporate explainability serves to avert blind reliance on engineer-level explanations, which may lack business context, and supports the open, pluralistic, and critical integration of AI into the institutional culture.

Whether this explanation should arise from the risk-management framework or be prepared independently is debatable; in either case, it must be clearly understandable at the corporate level.

Two documents are key: (i) *XAI-Corporation Fact Sheet*, a standardized document summarizing system parameters, averages, and key performance indicators, updated monthly; (ii) *XAI-Corporation Report*, intended for the Board of Directors and Senior Management, is issued quarterly. It must be strategic, concise, and business-oriented, eschewing technical jargon; its purpose is not to provide algorithmic traceability but to foster trust and ensure sound AI governance.

The structure should follow an executive format, ideally five paragraphs within a single page: (a) *executive summary* — purpose, KPI performance, and alignment with ethics, robustness, and legality; (b) *connection to business objectives*—translating technical metrics into financial impact, identifying value-adding features, and demonstrating explainability; (c) *system functionality*—principal variables, their decision logic, and behavior in critical scenarios; (d) *risk mitigation*—bias detection, drift identification, regulatory risk, auditability, and exception handling (always human-led); (e) *conclusion*—action-oriented findings, including whether the model is fully operational, expanding, or requiring resources for specific improvements

E/ *XAI-Client*

Clients require the most accessible and comprehensible explanations, rendered in natural language and, where deemed appropriate, bolstered by graphics, symbols, or multimedia. The purpose is to help them understand why a decision was made (e.g., loan denial), to apprise them of their right to request human review, and to enable the exercise of their right to appeal. The term *client* includes not only

customers with active relationships but also any person engaging with the bank, even through advertising.

The practical application of XAI for end users is the automated generation of customer reasoning statements (*Reasons Code*). These must explain why a decision was made, with special emphasis on the grounds for a rejection; they must articulate all causal factors (positive, negative, neutral), refrain from technical jargon and convoluted syntax, and avoid stereotyped or boilerplate formulations.

State-of-the-art explanation engines can yield concise, defensible reasons, such as: *Recent history of mortgage default or significant late payment (60+ days) / Debt-to-income ratio unacceptably high / Revolving credit utilization above 80% / Thin credit file lacking sufficient history.*

However, after a decade of increasing mechanization and depersonalization, European regulatory trends favor the re-humanization and individualization of customer service. Financial exclusion originates with a lack of understanding. Explanations must therefore offer customer-centric assistance, including counterfactual scenarios—what the client could do to obtain a different outcome in the future.

This transcends merely listing reasons: banks have a duty to contribute substantively to citizens' financial literacy. Every interaction should serve to educate, at least minimally, about banking functions.

XAI-Client statements must clearly apprise—preferably in bold if it is text; or by graphic narratives—of the right to human review, the broadest possible methods for invoking it, and the right to challenge the decision, specifying deadlines and the competent bodies (e.g., customer service, ordinary courts, market conduct regulators, independent financial customer authorities, data-protection or AI-oversight bodies, or mediation and arbitration entities).

The statement must also identify the bank's supervisory authority and the entity conducting the external audit of automated

explanations—if distinct from other AI auditors—and the date of the most recent *XAI-Engineer certificate of conformity*.

How *to Explain the Seemingly Inexplicable*

The necessity of justifying a computational decision to the average citizen clashes with the inherent opacity of the most advanced, highly predictive AI models.

The deployment of highly predictive models—virtually all of which are *Black Box*—is operationally untenable within the financial sector, not merely due to the increased burden of compliance with transparency requirements, but because their opacity hampers the early detection of coding errors and proxy-bias leakage. Such latent issues are often readily identifiable through direct inspection of code in *White-box* models, though this comes at the expense of lower predictive efficacy. Consequently, XAI functions not solely as a disclosure instrument but essentially as a mechanism for mitigating systemic risks inherent in high-performance AI.

Given the colossal data-processing throughput of advanced systems and the escalating complexity of risk vectors, a defining function of XAI is to ensure that interpretability does not constitute a performance constraint. This mandates a critical shift in industry practice, moving away from the mandatory inherent interpretability that characterized models under Basel II (e.g., linear regression) toward understandability secured through robust XAI governance.

As a result, a specialized subsector of the AI industry is rapidly crystallizing, dedicated to pioneering methods that permit humans to accurately interpret machine-generated decision processes, thus striving to reconcile explanatory transparency with maximal predictive performance.

IV

XAI TECHNIQUES

We are confronted with a heterogeneous landscape of solutions proposed by academics and computational practitioners—an expanding collection of techniques that introduce new ideas or, more often, partially overlap with existing methods by refining components, combining multiple approaches, or producing hybrid models.

More than four hundred techniques have been catalogued to date. Not all are suitable for every industrial setting; nonetheless, the breadth of this repertoire is significant. Irrespective of mainstream technological preferences, even the least-used techniques may eventually become essential for explaining specific artifiligent applications in particular business domains.

Efforts to systematize this diversity[72]—whether intended to produce a robust taxonomic tool[73] or an analytically pedagogical framework[74]—pursue shared objectives: model transparency, fidelity, stability, and the exhaustiveness of the *explicandum*. These systematization strategies converge on a dual classification of XAI

[72] Arrieta, Alejandro; Et al. (2019). "*Explainable artificial intelligence (XAI): Concepts, Taxonomies, Opportunities and challenges toward responsible AI*". 11. https://doi.org/10.1016/j.inffus.2019.12.012

[73] Vilone, Giulia; Luca Longo (2020). "*Explainable Artificial Intelligence: A Systematic Review*". Journal of School of Computer Science, College of Science and Health, Technological University Dublin, Dublin, Republic of Ireland. https://arxiv.org/pdf/2006.00093

[74] Molnar, Christof (2022). "*Interpretable Machine Learning: A guide for making Black Box models explainable*". 3rd. Ed. https://christophm.github.io/interpretable-ml-book/

techniques: (a) their scope (global vs. local) and (b) their relationship to the underlying model (model-specific vs. model-agnostic).

Agnostic Cicerones

Black-Box models—ML and DNN systems—operate through relationships between inputs and outputs that remain opaque to the average observer. Their principal advantage lies in their superior predictive capacity, making them particularly suitable for domains in which aggregate accuracy is paramount, such as fraud detection. Yet their key limitation is low interpretability.

A *Random Forest*, for example, aggregates thousands of decision trees, rendering the pathway to any given prediction exceedingly difficult to understand. This opacity presents considerable operational and compliance risks, particularly in the banking sector, which must detect and prevent deviations, illegal activities, and operational inconsistencies.

To combine the predictive power of *Black Box* models with the traceability and explainability required by regulated industries, a range of techniques can be applied after a model has been trained and has issued prediction—the so-called *post-hoc* techniques. Their purpose is to generate global or local interpretations of the model's behavior.

Model-agnostic techniques (those for which the architecture of the underlying model is irrelevant) can be applied to any ML system. They generally work by perturbing input data and measuring the corresponding effect on the output, allowing analysts to infer both the global decision logic (e.g., the relative importance of variables) and the reasoning behind specific individual predictions.

One drawback of *post-hoc* model-agnostic analysis is that the explanations generated may not reflect the model's internal logic. In some cases, XAI may introduce its own biases or distortions, producing explanations that reinterpret rather than faithfully capture

the model's reasoning. This exposes organizations to residual legal risk: an inaccurate explanation may unintentionally validate or obscure a biased decision.

A/ Post-hoc Global

A set of XAI techniques provides insight into the overall behavior of a *Black Box* model. *Permutation Feature Importance* (PFI) measures the relevance of a feature by quantifying the increase in model error when that feature's values are randomly permuted—thereby severing its relationship with the target. A feature is considered important if shuffling its values significantly degrades model performance[75].

Already in use within banking operations are *Partial Dependence Plots* (PDP), often applied as an explainability layer in credit scoring models using, for example, *XGBoost* (a non-additive tree model). PDPs visualize the influence of a variable on the model's output while marginalizing over all other variables. They show how predictions vary according to one or two independent variables, enabling assessment of the marginal effects of predictors and the nature of their relationship with the dependent variable.

PDPs display the average variation in predictions along a curve, obtained by modifying the value of a feature across all observations in the dataset and computing the resulting average impact. A key variant is *Individual Conditional Expectation* (ICE), which traces how predictions for each observation vary when one feature is changed while all others remain constant[76].

---

[75] Fisher, Aaron; Et al. (2019). "*All Models Are Wrong but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously*" Journal *of Machine Learning Research: JMLR* 20:
177. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8323609/.

[76] Greenwell, Brandon M. (2017). "*pdp: An R package for constructing Partial Dependence Plots*". The R Journal Vol. 9/1, June 2017. https://journal.r-project.org/articles/RJ-2017-016/RJ-2017-016.pdf

Accumulated Local Effects (ALE) address a major limitation of PDPs: the assumption of feature independence[77]. ALE plots assess the relationship between feature values and target variables while correcting for dependencies. *Feature Interaction* methods further decompose predictions into baseline terms, individual feature contributions, and an interaction component—capturing the Aristotelian notion that the whole may be greater than the sum of its parts.

Leave-One-Feature-Out (LOFO) approach evaluates feature importance by retraining the model without a given feature and comparing the resulting performance with that of the original model. If predictive performance deteriorates significantly when a feature is removed, that feature is deemed important.

Additional global techniques include *Functional Decomposition*, as well as *Prototypes and Criticisms*. A *prototype* is a representative input instance, whereas a *criticism* is an instance poorly represented by the prototypes, signaling anomalies, edge cases, or outliers not captured by the representative set.

Among global techniques, *Surrogate Models* deserve particular attention[78]. A global surrogate is an interpretable model trained to approximate the predictions of a *Black Box* model. By interpreting the surrogate, one indirectly interprets the underlying system. Surrogate models—also known as *emulators*, *approximation models*, *response-surface models*, or *metamodels*—seek to mimic the output of the opaque model as accurately as possible while remaining intrinsically interpretable.

---

[77] Apley, Daniel; Jingyu Zhu (2020). "*Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models*" Journal of the Royal Statistical Society Series B: Statistical Methodology 82 (4): 1059–86. https://arxiv.org/pdf/1612.08468

[78] Wilhelm, Alexander; Katharina A. Zweig (2024). "*Hacking a surrogate model approach to XAI*". https://arxiv.org/pdf/2406.16626

B/ Post-hoc Local

These techniques aim to clarify the reasoning behind individual predictions; they are currently preferred by both academic researchers and practitioners in large corporations. Examples include *Breakdown* (also called *Additive Feature Importance*) and *LOcal Rule-based Explainer* (LORE), which relies on locally trained decision trees.

A central tool in this category is the *Ceteris Paribus Plot* (CPP)[79], which illustrates how changes in a single feature alter the prediction for a specific data point while keeping all other variables constant. The procedure consists of selecting an observation, systematically varying the feature of interest across its domain, holding the remaining inputs fixed, and visualizing how the prediction shifts. Beyond its intrinsic explanatory value, CPP is foundational for other interpretability methods, most notably *Partial Dependence Plots* (PDPs). CPP is a model-agnostic method that can be creatively combined across many instances to perform higher-order analyses, such as model comparison, feature effect benchmarking, or the study of multiclass classification systems.

Closely related are the aforementioned *Individual Conditional Expectation* (ICE) plots, which disaggregate the PDP by displaying the effect of a feature on every individual observation. Each line represents how the model's prediction for a particular instance change as the feature varies. Whereas a CPP examines a single curve, an ICE plot overlays the curves for many or all instances, making heterogeneity and interaction effects directly observable. Diverging

[79] Kuźba, Michal; Et al. (2019). "*pyCeterisParibus: explaining Machine Learning models with Ceteris Paribus Profiles in Python*". Journal of Open-Source Software, 4(37), 1389. https://www.theoj.org/joss-papers/joss.01389/10.21105.joss.01389.pdf

slopes or shapes thus highlight variations in sensitivity that average-level metrics conceal[80].

One of the most influential techniques in modern XAI is *Local Interpretable Model-agnostic Explanations* (LIME). LIME explains a specific prediction by approximating the *Black Box* model in the local neighborhood of an instance with a simple surrogate model—often linear regression or a small decision tree. The algorithm perturbs the original input, generates synthetic data points, obtains model predictions for them, and then fits an interpretable model weighted by proximity to the original point. The quality of the explanation is generally assessed through the local fit coefficient. LIME is especially valued for its conceptual skepticism, ease of implementation, and efficiency in high-dimensional settings.

Another core technique, rooted in causal reasoning, is the *Counterfactual Explanation*[81]. Counterfactuals answer questions of the form: *What minimal change in the input would have produced a different outcome?* They emulate the human capacity to reason about alternatives and are particularly useful in regulated environments; for example, explaining to a rejected loan applicant which modifiable attributes would need to change for approval.

Scoped Rules, or *Anchors*, constitute a rule-based local method that identifies *if-else* patterns that reliably fix the model's behavior. Starting from an observed prediction, the technique perturbs the instance and searches for rules that, when satisfied, ensure the prediction remains stable. Variables not included in the rule should not affect the outcome, which offers a compact and precise explanation.

---

[80] Goldstein, Alex; Et al. (2015). "*Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation*". Journal of Computational and Graphical Statistics, 24(1):44–65, 2015. https://arxiv.org/pdf/1309.6392

[81] Wachter, Sandra; Et al. (2018). "*Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*". *Harvard Journal of Law and Technology* 31(2): 841–87. https://arxiv.org/pdf/1711.00399

Game-theoretic methods have also shaped *post-hoc* local explainability, most prominently through *Shapley Values*. In this framework, a model's prediction is seen as a cooperative game in which features are players contributing to the final payout. *Shapley Values* quantify the fair contribution of each feature by averaging its marginal impact across all possible coalitions. This produces a decomposition of the prediction into a baseline value plus a sum of feature attributions.

Building on this foundation, *SHapley Additive exPlanations* (SHAP) has become the most widely adopted XAI technique in business practice. SHAP unifies several interpretability methods into a mathematically consistent additive model. It computes the marginal contribution of each feature to the final prediction, assigning a "decisiveness coefficient" that ensures fairness and consistency across instances. The feature contributions, when summed with the baseline (mean prediction), exactly reproduce the model's output, which provides both local and global interpretability[82].

Insider Ushers

Insiders refer to XAI methods that are model-specific and rely on internal gradients, activations, architectural structure, or other model-specific components. They provide explanations that are tightly coupled with the functioning of the underlying model, especially within Deep Learning systems. These methods fall into two categories: *model-specific XAI techniques* and *inherently interpretable models*.

Among the model-specific techniques are frameworks such as *DeepExplain, Deep Visualization*, *INNvestigate*, *RISE*, *TCAV*, *tf-*

[82] Lundberg, Scott; Su-In Lee (2017). "*A unified approach to interpreting model predictions*". In Advances in Neural Information Processing Systems, pages 4765–4774, Long Beach, California, USA, 2017. Neural Information Processing Systems Foundation, Inc. https://arxiv.org/pdf/1705.07874

Explain, *Integrated Gradients*, and *Rationale*. A prominent method is *Deep Learning Important FeaTures* (DeepLIFT), which compares the activation of each neuron to a reference activation to determine its precise contribution to the final output. This allows for the construction of a traceable pathway linking input perturbations to predictions across all layers.

Saliency Maps are among the most influential tools for interpreting neural networks[83]. They identify the input regions—image pixels, text tokens, or other components—that most strongly affect the model's output. The method computes the sensitivity of the prediction to small changes in each input feature. Features whose slight perturbation produces large changes in the prediction are considered salient. The result is typically visualized as a heatmap, highlighting the critical input regions the model relied upon.

A more sophisticated technique is *Gradient-weighted Class Activation Mapping* (Grad-CAM), which produces class-specific heatmaps for convolutional neural networks (CNNs). The process consists of: *Forward pass*—the input image is processed through the CNN to produce feature maps—; *Backward pass*—the gradient of the target class score is computed with respect to the final convolutional layer—; *Weight calculation*—the gradients are averaged to determine the importance of each feature map—; *Class activation map construction*—a rectified, weighted sum of the feature maps yields a low-resolution heatmap—; and *Visualization*—the heatmap is upscaled and superimposed on the input image, with higher-activation regions indicating areas critical to the classification—.

[83] He, Sen: Nicolas Pugeault (2017). "*Deep saliency: What is learnt by a deep network about saliency?*". International Conference on Machine Learning-Workshop on Visualization for Deep Learning, pages 1–5, Sydney, Australia, 2017. ICML. https://arxiv.org/pdf/1801.04261

For systematization, it is appropriate to include techniques often treated separately in computational literature, particularly those focused on explaining DNN and Deep Learning systems.

Learned Features acknowledges that DNN automatically extracts high-level abstractions through its hidden layers. Early layers identify simple edges and textures; intermediate layers detect patterns and shapes; and deeper layers recognize complete objects. *Feature Visualization* makes these learned abstractions visible, facilitating intuitive insight into the internal representations of the network.

Detecting Concepts addresses two limitations of many feature-attribution methods: (i) low interpretability of primitive features (e.g., individual pixels), and (ii) limited expressiveness due to the finite number of input features. This approach maps user-defined high-level concepts into the network's latent space and identifies where these concepts manifest within the model's learned internal representations. It effectively translates complex transformations in hidden layers into human-understandable abstractions[84].

Adversarial Examples introduces small, strategically designed perturbations to an input to cause a misclassification. While related to counterfactuals, adversarial examples differ in intent: they are crafted to deceive the model, uncovering vulnerabilities and identifying decision boundaries with extreme sensitivity[85].

Finally, *Influential Instances* examines the impact of individual training samples on model predictions. The method identifies influential or outlier points, allows debugging of mislabeled or problematic instances, and supports model retraining if necessary. A practical variant, *Influence Functions*, estimates how removing or altering an instance would affect the model without requiring full

[84] Poeta, Eleonora; Et al. (2023). "*Concept-based Explainable Artificial Intelligence: A Survey*". https://arxiv.org/pdf/2312.12936

[85] Baniecki, Hubert; Przemyslaw Biecec (2025). *"Adversarial attacks and defenses in explainable artificial intelligence: A survey"* https://arxiv.org/html/2306.06123v4

retraining, using derivative-based approximations to reduce computational cost.

White Box

This model-specific category of XAI refers to systems that, strictly speaking, should not be classified as explainability techniques, since they consist of *transparent models*—also known as *White Box*, *ante-hoc*, *interpretable-by-design*, or *self-explainable* models. Their internal mechanics are inherently legible because interpretability is embedded directly into their architecture.

Their main advantage lies in their structural transparency, which enables human observers to trace decision vectors and understand the relationships between inputs, transformations, and outputs with relative ease. Direct inspection of the code allows developers to verify the underlying logic, detect errors, and ensure reliability at a granular level. In highly regulated environments such as banking, this traceability has long been a fundamental requirement, particularly under frameworks derived from Basel II.

However, transparent models generally possess significantly lower predictive power compared with modern ML architecture. Their commitment to interpretability imposes constraints on complexity, limiting their ability to capture intricate nonlinear relationships in risk-related data. Moreover, their development often requires the costly design of bespoke and highly tailored architecture.

Inherently interpretable models include classic approaches such as linear and logistic regressions, decision trees, and rule-based systems, as well as less frequently referenced models such as *Generalized Linear Models* (GLM), *Generalized Additive Models* (GAM), *Support Vector Machines* (SVM**s**), *RuleFit*, and *Naïve Bayes*.

In these systems, the output reflects both the model's architecture and the structure of the inputs. The *Bayesian Case Model* (BCM), for

instance, identifies prototypes that best represent the clusters in a dataset through simultaneous inference of cluster labels, prototypes, and relevant features[86].

Gaussian Process Regression (GPR) is a non-parametric technique that does not assume a specific functional form for the estimator. It is robust to missing data and interpretable because the resulting features weights provide an explicit measure of relevance[87]. *Generalized Additive Models* (GAMs) combine linear predictors with shape functions trained on one or two variables, enabling users to visualize and understand the contribution of each feature through intuitive bar or line plots[88].

Other interpretable models referenced in the computational literature include *Oblique Treed Sparse Additive Models* (OT-SpaMs), *Transparent Generalized Additive Model Trees* (TGAMT), *Multi-Run Subtree Encapsulation*, *Probabilistic Sentential Decision Diagrams* (PSDD), *Mind the Gap Models* (MGM), *Supersparse Linear Integer Models* (SLIM), unsupervised interpretable word-

[86] Kim, Been; Cynthia Rudin, and Julie A Shah (2014). "*The bayesian case model: A generative approach for case-based reasoning and prototype classification*". Advances in Neural Information Processing Systems, pages 1952–1960, Montreal, Quebec, Canada, 2014. Neural Information Processing Systems Foundation, Inc. https://arxiv.org/pdf/1503.01161

[87] Caywood, Matthew S.; Et al. (2017). "*Gaussian process regression for predictive but interpretable machine learning models: An example of predicting mental workload across tasks*" Frontiers in human neuroscience, 10:647–665, 2017. https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2016.00647/full

[88] Yin Lou; Et al. (2013) "*Accurate intelligible models with pairwise interactions*". In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 623–631, Chicago, Illinois, USA, 2013. ACM. doi: 10.1145/2487575.2487579 (previous submission to the authors).

sense disambiguation systems, feature-map visual explanations[89], and symbolic graph reasoning[90].

Complementary Strategies

Despite the diversity of contemporary XAI methods, none of them, when evaluated from the standpoint of managing High-Risk AI Systems (HRAIS), provides what could be considered a sufficiently robust or universally acceptable solution to the explainability challenge.

Worse still, under certain conditions, many of these methods may produce contradictory or biased explanations. This risk compounds other well-documented issues that undermine interpretability: limited reproducibility of results, training inconsistencies, unstable prediction behavior, biases embedded in input data, and the accuracy and impartiality of explanatory outputs.

In response, researchers are developing complementary approaches designed to improve interpretability, strengthen performance evaluation, and mitigate issues such as overfitting. These approaches include extracting information from intermediate layers of DNNs; aggregating interpretability metrics and developing adversarial models to quantify explainability; reducing the number of parameters to be optimized; and employing advanced visualization methods to facilitate human comprehension.

Additional strategies involve model simplification, restricting features to those with clear business meaning, analyzing data for bias

---

[89] Zintgraf, Luisa; Et. al. (2017). "*Visualizing deep neural network decisions: Prediction difference analysis*". 5th International Conference on Learning Representations, Toulon, France, 2017. ICLR. https://arxiv.org/pdf/1702.04595

[90] Liang, Xiaodan; Et al. (2018). "*Symbolic graph reasoning meets convolutions*". Advances in Neural Information Processing Systems, pages 1853–1863, Montreal, Canada, 2018. Neural Information Processing Systems Foundation, Inc. https://www.cs.cmu.edu/~epxing/papers/2018/Liang_etal_nips18.pdf

or lack of impartiality, and assessing reproducibility during model development and deployment. The overarching objective is to construct hybrid systems that seamlessly merge the most reliable explanatory methods.

A powerful approach within this paradigm is *Model Distillation*, which transfers the decision logic of a complex, opaque *teacher model*—often an intricate *Random Forest*—into a simpler, interpretable *student model*, such as a scorecard or a decision tree. The student model learns not only from the original inputs but also from the teacher's soft predictions and internal representations.

This technique allows institutions to retain much of the predictive performance of the *Black Box* model while simultaneously generating an interpretable surrogate suitable for regulatory justification. It substantially reduces the risk of infidelity, i.e., the danger that a *post-hoc* explanation contradicts the teacher model's true logic—and thus reinforces the institution's regulatory defense.

Nevertheless, the design of transparent models still faces considerable challenges. Interpretability often results from deliberate constraints on the optimization space—for example, limiting the number of variables used for prediction, reducing rule complexity, restricting decision-tree depth, or reducing neural-network width. These restrictions improve explainability but typically degrade accuracy.

There is an ongoing debate about establishing stricter criteria for what constitutes an inherently interpretable model. Proposed characteristics include: *additivity*, ensuring that input effects are separable and their contributions easily aggregated; *sparsity*, favoring concise explanations focused on the most relevant factors; *linearity*, ensuring proportional relationships between inputs and outputs; *monotonicity*, making increasing or decreasing influences predictable across ranges; *concept decoupling*, where neural networks preserve

conceptual separability across layers; *dimensionality reduction*, providing visual post-hoc tools for human interpretation.

Another essential area involves improving model governance to ensure fairness, eliminate biases in training data, incorporate expert judgement, and maintain performance and control frameworks. Error analysis, including the balance between bias and variance, remains essential. However, data-protection constraints can limit the detection of biases when sensitive attributes are not stored or cannot be processed.

A concept receiving particular attention is *counterfactual fairness*, which evaluates whether individuals with similar characteristics—differing only in protected attributes—receive similar model outcomes. This provides a principled, causal metric for assessing discrimination.

Finally, stakeholders increasingly emphasize the need for effective human oversight throughout all stages of Artifiligence operation. XAI tools must alert human decision-makers when an automated decision approaches risk or fairness thresholds, enabling timely intervention and ensuring that algorithmic processes remain under meaningful human control.

Critical Evaluators

Parallel to the development of explanatory techniques, the scientific community has underscored the absence of a comprehensive layer of validation. As a result, significant effort has been devoted to designing methodologies for evaluating XAI systems, particularly with respect to explanation fidelity and human usefulness.

This is an expanding field where few assumptions can be treated as definitive. Mastery of evaluation methods is essential for supervisory engineers seeking to define and standardize regulatory requirements, as well as for practitioners responsible for proposing improvements to XAI risk management.

Examples of formal evaluation include *objective* or *heuristic-based* methods. *Explain and Ime*, for instance, employ quantitative metrics such as filter interpretability and location instability. *InterpNet* proposes three evaluation metrics: BLEU, which measures sentence similarity using n-gram matching; METEOR, which incorporates semantic proximity via alignment of trained word embeddings; CIDEr, which evaluates neural-network-generated descriptions by comparing weighted n-grams to human references using TF–TF-IDF-based weighting[91]. Another approach is the method assessing the risk of *generating counterfactual instances* that are mere items of the classifier rather than grounded in plausible data distributions[92].

Human-centered or *user-based* evaluations can be qualitative or quantitative and may involve textual, visual, rule-based, or hybrid formats. They draw on diverse methodological traditions, from data clustering[93], ML techniques[94] to Naïve Bayes[95], autonomous agents,

---

[91] Barratt, Shane (2017). "*Interpnet: Neural introspection for interpretable deep learning*". In NIPS Symposium on Interpretable Machine Learning, pages 47–53, Long Beach, California, USA, 2017. NIPS. https://arxiv.org/pdf/1710.09511

[92] Laugel, Thibault; El al. (2019). "*The dangers of post-hoc interpretability: Unjustified counterfactual explanations*". In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, (IJCAI), pages 2801–2807, Macao, China, 2019. International Joint Conferences on Artificial Intelligence Organization. https://arxiv.org/pdf/1907.09294

[93] Kim, Been; Et al. (2015). "*Mind the gap: A generative approach to interpretable feature selection and extraction*". In Advances in Neural Information Processing Systems, pages 2260–2268, Montreal, Quebec, Canada, 2015. Neural Information Processing Systems Foundation, Inc.
https://proceedings.neurips.cc/paper_files/paper/2015/file/82965d4ed8150294d4330ace00821d77-Paper.pdf

[94] Krause, Josua; Et al. (2016). "*Interacting with predictions: Visual inspection of black-box machine learning models*". In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 5686–5697, San Jose, California, USA, 2016. ACM. https://chi2016-prospector.pdf

[95] Kulesza, Todd; Et al. (2011). "*Why-oriented end-user debugging of Naive Bayes text classification*". ACM Transactions on Interactive Intelligent Systems (TiiS), 1(1):2:1–2:31, 2011. https://openaccess.city.ac.uk/id/eprint/12414/10/TiiS-1129-1149.pdf

expert systems, SVMs, neural networks, case-based reasoning, Kernel methods, decision trees, and guideline-based classifiers.

A third line of inquiry focuses on *benchmarking XAI techniques*, comparing methods by examining their sensitivity to data perturbations[96], parameter customization[97], or explanation completeness[98]. In credit-scoring contexts, comparative analyses frequently contrast SHAP with PDP or, more commonly, SHAP with LIME. Metrics such as *K-means Silhouette* and *Spectral Clustering Silhouette* often suggest that LIME yields explanations with lower internal consistency—implying greater XAI risk—while SHAP generally demonstrates stronger discriminatory power and more coherent clustering behavior, resulting in better stability and regulatory defensibility[99].

Across these approaches, the consensus is growing that XAI evaluation—central to shaping the future relationship between humans and AI systems—requires collaboration among diverse research communities. This effort must culminate in a global forum capable of sustaining an open, critical, interdisciplinary debate about consciousness, language, interpretation, and other fundamental dimensions of human cognition now challenged by the paradigm shift brought about by Artifiligence[100].

---

[96] Adebayo, Julius; Et al. (2018). "*Local explanation methods for deep neural networks lack sensitivity to parameter values*". In 6th International Conference on Learning Representations, Vancouver, Canada, 2018. ICLR. https://openreview.net/pdf?id=SJOYTK1vM

[97] Alvarez-Melis, David; Tommi S. Jaakkola. (2018). "*On the robustness of interpretability methods*". In Proceedings of the 2018 ICML Workshop in Human Interpretability in Machine Learning, pages 66–71, Stockholm, Sweden, 2018. ICML. https://arxiv.org/pdf/1806.08049

[98] Gevrey, Muriel; Et al. (2003). "*Review and comparison of methods to study the contribution of variables in artificial neural network models*". Ecological modelling, 160(3):249–264, 2003. https://www.sciencedirect.com/science/article/pii/S0304380002002570?via%3Dihub

[99] Salih, Ahmed; Et al. (2024). "*A Perspective of Artificial Intelligence Explainable methods of Explanation: LIME vs SHAP*". https://arxiv.org/html/2305.02012v3

[100] Longo, Luca; Et al. (2024). "*Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions*" Information Fusion Volume 106, June 2024, 102301
https://www.sciencedirect.com/science/article/pii/S1566253524000794

V

EXPLANATION RISK MANAGEMENT

Each risk entails a distinct management framework that must be addressed through specific methodologies designed to measure and control the bank's exposure to that source of vulnerability.

Credit risk, for instance—defined as the possibility of incurring losses due to borrowers or counterparties failing to meet their payment obligations—may be assessed through ratings, scorings, capital structure valuations (e.g., the Merton model), credit portfolio loss distributions (e.g., CreditRisk+), or calculations concerning the return on required capital (e.g., the *Standardised Approach* or the *IRB approach*[101]).

Market risk, which seeks to quantify potential losses stemming from adverse movements in market variables (interest rates, exchange rates, equity prices), employs a spectrum of methodologies to measure *Value at Risk* (VaR): variance-covariance (parametric VaR), scenario-based approaches (historical simulation VaR, *Monte Carlo VaR*), and conditional loss metrics (*Tail VaR* or *Expected Shortfall*).

As new categories of risk (and mutations of traditional ones) emerge, cross-cutting or strategic risk management models are being developed. This is visible in crisis management, where stress-testing models simulate the impact of extreme yet plausible economic and financial scenarios on the bank's solvency and liquidity; in internal

[101] European Banking Authority - EBA (2023). "*Machine learning for IRB models: Follow-up report from the consultation on the discussion paper on machine learning for IRB models*". https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf

capital planning and risk appetite frameworks; and in the integration of sustainability considerations through the management of *Environmental, Social, and Governance* (ESG) risks, including the quantification of transition risk.

At first glance, the expansion of Artifiligence creates a vast new risk landscape—raising the question of whether each AI-related risk should be integrated into the traditional risk category it resembles, or whether the potentially ubiquitous and systemic nature of AI risk justifies the establishment of a dedicated, transversal *All-over AI Risk Framework*.

GPAI, for example, introduces inherent technological risks (data privacy, embedded bias), amplifies known cybersecurity threats, gives rise to unprecedented forms of systemic risk, and generates diverse risks linked to system performance, such as robustness, synthetic data handling, and explainability[102].

Explainable AI must be embedded within the bank's overall risk management system, as the opacity characteristic of *Black Box* models can provoke widespread process failures, hinder auditability, harm customers, and erode the institution's reputation. A failure of XAI can trigger cascading adverse effects.

Explainability is essential for demonstrating compliance with laws and regulations. It enables verification that decisions are transparent, unbiased, grounded in legitimate criteria, and non-discriminatory. Moreover, if a bank cannot adequately explain decisions perceived by customers as arbitrary, unfair, or erroneous, it undermines trust, damages its brand, and produces significant business losses involving clients and investors.

Under the current regulatory framework, XAI risk is treated as a form of *model risk* (specifically, *secondary model risk*), managed

[102] Shabsigh, Ghiath; El Bachir Boukherouaa (2023). "*Generative Artificial Intelligence in Finance: Risk Considerations*". IMF Fintech Notes, 2023/006, International Monetary Fund. https://doi.org/10.5089/9798400251092.063

under the broader umbrella of *operational risk*. An inability to understand how a model reaches its decisions constitutes a deficiency in internal control or a process failure. This may stem from design specification errors (unrealistic assumptions; omission of relevant variables; methodological limitations), implementation errors (faulty programming; incorrect integration into legacy systems), or usage errors (misapplication; inadequate inputs).

At present, two regulatory vectors primarily shape the management of *XAI Model Risk* (XAI MRM): Basel III and the DORA Regulation[103].

A/ Basel III

Basel III is an international regulatory framework composed of multiple documents that establish minimum capital, liquidity, and risk management requirements for banking institutions[104]. Among its innovations are the *Liquidity Coverage Ratio* (LCR) and the *Net Stable Funding Ratio* (NSFR), which address liquidity risk—a major weakness in earlier frameworks.

Basel III risk models determine the calculation of risk-weighted assets (RWA), which define the minimum capital banks must hold. The framework is structured around three pillars—minimum capital requirements, supervisory review, and market discipline—and governs three major risk types.

For credit risk, the *Standardised Approach* (SA) assigns regulator-defined risk weights to exposures, whereas the *Internal*

[103] European Commission (2022). *Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014 and (EU) 2016/1011*. DORA https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2554

[104] Basel Committee on Banking Supervision – BCBS (2017). "*Basel III: Finalising post-crisis reforms*". https://www.bis.org/bcbs/publ/d424.pdf

Ratings-Based (IRB) approach allows banks to use internal models. IRB includes two variants: *Foundation IRB* (F-IRB), where institutions estimate the Probability of Default (PD) and rely on regulatory values for other parameters (e.g., LGD); and *Advanced IRB* (A-IRB), where banks internally estimate all risk parameters.

For market risk, institutions may use the *revised Standardised Approach*—more risk-sensitive than earlier versions—or internal models, which remain permitted but subject to more stringent validation.

Regarding operational risk, Basel III replaced previous methods with a *simplified Standardised Approach*, eliminating the *Advanced Measurement Approach* (AMA) due to deficiencies revealed during the financial crisis.

Beyond capital measurement, qualitative AI risk management requires a set of action vectors whose systematization depends on each bank's corporate tradition. The *NIST AI Risk Management Framework* identifies four essential functions: *Govern*, *Map*, *Measure*, and *Manage*[105] .

In European *Model Risk Management* practice, it is common to distinguish: (a) *Risk and Control Self-Assessment* (RCSA) with Risk Mapping grouping operational risks by event category (fraud, system failures, human error) and by business unit;

[105] NIST Op.cit. "*Fundamental Functions of NIST AI ARF.— At its core, the NIST AI RMF is built on four functions: Govern, Map, Measure, and Manage. These functions are not discrete steps but interconnected processes designed to be implemented iteratively throughout an AI system's lifecycle. The 'Govern' function emphasizes the cultivation of a risk-aware organizational culture, recognizing that effective AI risk management begins with leadership commitment and clear governance structures. 'Map' focuses on contextualizing AI systems within their broader operational environment, encouraging organizations to identify potential impacts across technical, social, and ethical dimensions. The 'Measure' function delves into the nuanced task of risk assessment, promoting both quantitative and qualitative approaches to understand the likelihood and potential consequences of AI-related risks. Finally, 'Manage' addresses the critical step of risk response, guiding organizations in prioritizing and addressing identified risks through a combination of technical controls and procedural safeguards.*"

(b) *Key Risk Indicators* (KRI) monitoring, offering early warnings of increased risk and maintaining internal loss databases; (c) *mitigation and response mechanisms*, including internal controls, policies, procedures, segregation of duties, *Business Continuity Plans* (BCP), and *Disaster Recovery Plans* (DRP); and (d) *Governance*, pervading all business processes until it becomes organizational culture, with clear role and responsibility definitions[106].

This framework requires the production of extensive documentation to ensure transparency of the bank's risk profile[107]. Reports may be grouped by audience (internal or external) and purpose (governance, compliance, or monitoring), and evolve with regulatory changes, cybersecurity developments, technological shifts, macroeconomic instability (inflation, interest rates, tariffs), and sustainability or transition-related exposures[108].

High-level governance documents include: (i) *Risk Appetite Framework*, defining the amount and type of risk the bank is willing to assume; (ii) *Global Risk Management Policy*, outlining governance structures, roles, responsibilities, procedures, and methodologies[109]; and (iii) *Statement of Specific Risks*, detailing how each material risk type is handled.

---

[106] Basel Committee on Banking Supervision – BCBS (2019). "*High-level summary of Basel III reforms*". https://www.bis.org/bcbs/publ/d424_hlsummary.pdf

[107] Basel Committee on Baking Supervision – BCBS (2013). "*Principles for effective risk data aggregation and risk reporting*". https://www.bis.org/publ/bcbs239.pdf

[108] European Central Bank – ECB (2024). "*Guide on effective data aggregation and risk reporting*". https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guides240503_riskreporting.en.pdf

[109] Arndorfer, Isabella; Andrea Minto (2015). "*The "Four Lines Defense Model" for Financial Institutions*". Financial Stability Institute – Bank for International Settlements BIS. https://www.bis.org/fsi/fsipapers11.pdf

Internal monitoring reports include[110]: (iv) *Line Risk Reports*, showing daily or weekly exposure relative to limits[111]; (v) *Capital and Solvency Reports*, describing capital adequacy and risk coverage; (vi) *Stress Test & Scenario Analysis Reports*; and (vii) *Operational Risk Reports*, documenting loss events, near misses, incidents, and corrective actions.

Regulatory reports, designed for supervisory compliance and market transparency, include: (viii) *Annual Management Report / Risk Section*; (ix) *Pillar 3* disclosures; (x) *Annual Corporate Governance Report*; and (xi) *Report to the Risks Center*, informing the supervisor of all credit exposures.

The extensive integration of Artificial Intelligence into banking operations—and the pervasive impact of its intrinsic risks—forces a re-evaluation of the adjustments required in each of these documents. Ultimately, this process may lead to the creation of a new global reporting framework centered on the omnipresence of AI.

B/ DORA

The *Digital Operational Resilience Act* (DORA) *Regulation* addresses the operational risks arising from the growing dependence on Information and Communication Technologies (ICT), a framework necessarily applicable to AI systems.

Regarding the risk of losses resulting from inadequate or failed internal processes, personnel, systems, or external events, DORA elevates its management to a strategic level by focusing on: (a)

[110] Agarwal, Ruchi; Sangay Kallapur (2019). "*Four ways to improve risk reporting*". California Management Review 63 (4).https://doi.org/10.1177/00081256211019801

[111] Many software applications bring incorporated charts, dashboards and abundant infographic that can complement the report; typically, *Risk Tolerance, High-Velocity Risks Regulatory Obligations*, *Audit Plan*, *Internal Audit Findings Summary*, *Risk Incident Management*, *Centralized Control Reporting*, *Program-Wide Issue Summary*, *Objectives & Strategy Report.*

Systems risk, including failures or interruptions in the bank's hardware and software; (b) *Cybersecurity risk*, involving external threats that jeopardize the integrity, availability, or confidentiality of information and systems; and (c) *Outsourcing* or *Third-Party Risk* (TPR), derived from dependence on external ICT service providers whose failures may paralyze critical banking functions.

With respect to *outsourcing risk*, DORA expands supervision to ICT service providers and requires banks to: develop a reasoned and traceable strategy for all third-party relationships; maintain an up-to-date and complete register of all contractual agreements for ICT services, especially for those supporting critical functions; and actively assess, prevent, and manage concentration risk arising from dependence on a limited number of pivotal providers. Contracts with third parties must include specific mandatory clauses such as access and audit rights for authorities, continuity safeguards, and exit strategies.

By establishing a single regulatory framework for digital resilience across the European Union, DORA obliges entities to prevent, withstand, respond to, and recover from ICT disruptions, thereby mitigating their potential impact on financial stability and customer service.

Under DORA, the governing body retains ultimate responsibility for ICT risk management and must: define, approve, and regularly monitor the digital operational resilience strategy; establish a comprehensive and documented ICT risk management framework aligned with the institution's risk appetite; continuously map and classify all ICT assets, functions, systems, and dependencies—especially those supporting critical functions; and implement security and protective measures to minimize incident impact.

DORA harmonizes the response to technological failures through systematic detection and registration of ICT incidents, classification of incidents according to harmonized severity and impact criteria, timely notification to competent authorities, and the obligation to

inform customers about incidents with potential negative consequences for their financial interests.

Furthermore, DORA requires banks to prove that their systems can withstand attacks by performing annual comprehensive testing of all critical ICT systems and applications, including vulnerability analyses and advanced tests such as *Threat-Led Penetration Testing* (TLPT). Systemically important entities must undergo TLPT based on real threat scenarios at least every three years.

DORA also promotes information and intelligence sharing on cyber threats, encouraging institutions to form cyber threat intelligence agreements to enhance collective awareness and strengthen the sector's defensive capacity.

XAI Governance

The banking sector stands at a critical juncture where the unavoidable need to adopt high-performance AI models is only feasible within a stringent framework of explainability. XAI is not an optional feature but a structural requirement of the new technological architecture.

To ensure effective, sustainable, and continuously compliant XAI, organizations must integrate explainability from the earliest stages of AI implementation. The prevailing trend advocates for embedding XAI into *Data Protection by Design* methodologies (Art. 25 GDPR) and adopting algorithmic audits as standard practice. This ensures that explainability is proactive rather than reactive, reinforcing accountability and human oversight in automated decision-making.

These audits verify the technical implementation of XAI and ensure that its integration across business areas complies with principles of fairness, transparency, and accountability. They complement XAI's technical capabilities with systemic oversight. Risk must be prioritized by refining *Impact–Likelihood Matrices*, *Key Risk Indicators*, and alignment with the institution's *Risk Appetite*.

High-quality documentation and a clear allocation of responsibilities are critical. Explicit functions must be assigned, and procedures established for accountability within a continuous-improvement process based on feedback learning—an ideal in which system performance and compliance verification converge (*compliance by performance*).

The importance of interdisciplinarity must be emphasized. The core challenge in XAI lies in translating the model's technical reasoning into a narrative that is intelligible to humans. To achieve this, bank teams must include not only engineers and mathematicians but also experts from the humanities, particularly linguistics, ethics, and law. Their contribution is essential for anticipatory governance and ensures that technical explanations are both legally defensible and socially intelligible, transforming algorithmic metrics into meaningful human reasons.

Among all governance considerations, two areas demand special attention: Data and Bias.

A/ Not just any Data

Mastery of data will define the successful bank. The full benefits of Artifiligence cannot be realized unless corporations evolve into sophisticated data laboratories[112].

Banks must deeply understand the data they handle—its utility and its limitations—expanding on current methodologies such as *Data Sheets for Datasets*[113]. In decision-centric sectors, data

[112] Provost, Foster; Tom Fawcett (2013). "*Data Science for Business: What You Need to Know about Data Mining and Data-Analytic*". O'Reilly Media.

[113] In addition to the regulations already mentioned, particular attention must be paid, among others, to the following instruments within the so-called *EU Digital Acquis (EU Digital Rulebook)*: *Regulation (EU) 2022/2065 (Digital Services Act – DSA); Regulation (EU) 2022/1925 (Digital Markets Act - DMA); Regulation (EU) 2023/2854 (Data Act); Regulation (EU) 2018/1807 (Free Flow of Non-Personal Data Regulation); Directive (EU) 2019/1024 (Open Data Directive); Directive (EU) 2022/2555 (NIS2 Directive); Regulation (EU)*

optimization is essential to operational efficiency. Interaction with all data types—not only text but any format—must be seamless.

Some categories, such as structured data, have accompanied the entire Digital Revolution[114]. Others remain less familiar—such as *primitive data*, the most basic and indivisible units used across programming languages—yet they are essential to understanding and monitoring AI systems[115].

Furthermore, a distinction exists between data organizations that are legally obliged to collect (e.g., traditional credit data) and data whose storage and processing are prohibited by law (e.g., certain forms of alternative data), despite their uncontrolled circulation and widespread use in ML and DNN training contexts.

Synthetic data is particularly relevant, having already proven valuable in banking areas such as fraud detection[116]. This artificial data is generated through mathematical models or ML algorithms and can be ameliorated with *Privacy-Enhancing Technologies* (PETs) using statistical methods (Bayesian networks, conditional copulas, tree-based sequential synthesizers) or deep generative methods (GANs, Transformers, and LLMs).

2024/2847 (Cyber Resilience Act - CRA); Regulation (EU) 2024/1183 (European Digital Identity - Revised eIDAS Regulation); Regulation (EU) 910/2014 (eIDAS Regulation, original); Directive (EU) 2018/1972 (European Electronic Communications Code - EECC); Regulation (EU) 531/2012 (Roaming Regulation); Directive (EU) 2019/790 (Copyright in the Digital Single Market Directive); Directive (EU) 2010/13 (Audiovisual Media Services Directive - AVMSD); Regulation (EU) 2019/1150 (Platform-to-Business Regulation - P2B); Regulation (EU) 2017/2394 (Consumer Protection Cooperation Regulation); Directive (EU) 2016/2102 (Web Accessibility Directive).

[114] Grus, Joel (2019). "*Data Science from Scratch (First Principles with Python)*". 2nd Edition. O'Reilly Media.

[115] Hastie, Trevor; Robert Tishirani and Jerome Friedman (2008). "*The Elements of Statistical Learning (Data mining, Inference, Prediction*)". Springer. 2nd. Edition https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf

[116] Personal Data Protection Commission - PDPD [Singapore] (2024). "*Privacy Enhancing Technology (PET): Proposed Guide on Synthetic Data Generation V.1 (2024) - PDPC Singapore* https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/other-guides/proposed-guide-on-synthetic-data-generation.pdf

Data obfuscation is a PET technique applicable to anonymization, pseudonymization, secure storage, and software testing. The PET domain also encompasses: synthetic data generation for privacy-preserving machine learning; differential privacy for expanding research opportunities; and zero-knowledge proofs to verify information without disclosure; encrypted data processing is another critical PET domain, covering homomorphic encryption (enabling computation on private, undisclosed data), multi-party computation, and trusted execution environments; federated analytics also plays a key role in privacy-preserving machine learning, though distributed learning introduces re-identification risks through singularization, vulnerability detection, or inference attacks.

These measures are essential but operate independently of the obligation to mitigate risks associated with data privacy leaks caused intentionally or negligently by data controllers, careless behavior by data subjects, criminal exploitation of personal data, cybersecurity threats, and poor-quality inputs. For instance, common strategies to reduce hallucinations in LLMs include prompt engineering (one-shot, few-shot, chain-of-thought); *Retrieval-Augmented Generation* (RAG) utilizing vector databases, chunking, embedding techniques, and retrieval components; or, as a more costly alternative, model fine-tuning.

B/ The Fight Against Bias

A fundamental component of XAI is the fight against bias. This proclivity often originates in the data, is then assumed and amplified by the model, and generates severe consequences for equality, individual freedoms, and the democratic order[117]. Understanding the sources and typologies of bias is key to using XAI for detection and mitigation.

[117] O'Neil, Cathy (2016). "*Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*". New York: Crown Publishing Group.

Historical bias reflects pre-existing societal prejudices embedded in the data, causing the model to reproduce and reinforce inequalities. *Sampling bias* arises when training data does not accurately reflect the population the model is intended to serve. *Measurement bias* results from poor, inconsistent, or unequal data collection or categorization practices.

Labeling bias stems from subjective or inconsistent annotations (stigma, prejudice, cancellation), often unconscious, which the model subsequently learns. *Algorithmic bias* is introduced during the model's design and training, where optimization objectives may favor majority groups at the expense of minorities.

Confirmation bias occurs when the model disproportionately reinforces established trends, validating existing correlations and hindering the detection of inequitable patterns. *Intersectional bias* amplifies discrimination when multiple protected attributes intersect (e.g., gender plus race)[118].

Preventing bias is essential under EU regulation and in soft-law jurisdictions. For instance, in the UK, explanations must include fairness assessments, specifying whether outputs have discriminatory effects and whether formal fairness metrics were integrated into the system's design, consistent with the *Equality Act 2010*[119]. This means explainability must detail not only *how* a decision was made, but also *why* it is not discriminatory.

Ethics in AI transcends mere technical mitigation. It requires proactive governance to identify potentially illegal or inequitable scenarios before systems are created or deployed, translating ethical principles into verifiable models and ensuring that technology serves the collective welfare.

---

[118] Buolamwini, Joy; Timnit Gebru (2018). "*Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification".* Proceedings of Machine Learning Research, 81, 77-91. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[119] The National Archive [UK] "*Equality Act 2010*". https://www.legislation.gov.uk/ukpga/2010/15/contents/enacted

Mitigation strategies span all stages of data processing: increasing training data and applying resampling techniques to balance underrepresented groups; adjusting optimization functions to incorporate fairness penalties (e.g., *Demographic Parity*, *Equalized Odds*); and continuously monitoring explanations, since explanatory drift may occur when proxy features are used or decisions are justified indirectly.

Ultimately, XAI is a *sine qua non* in the fight against bias. It transforms equity management from reactive mitigation to a proactive, system-level process, fostering not only more equitable and compliant systems but also institutions whose decisions are transparent, defensible, and aligned with democratic values.

.

A Checklist Draft

Not all areas of a bank will deploy High-Risk AI Systems (HRAIS). However, an internationally operating bank of systemic importance will likely undertake its own GPAI developments. This transforms its role from a mere *deployer* to a *downstream modifier*, thereby increasing responsibility across all lines of activity.

Accordingly, a key function of XAI Governance, aligned with a clear delineation of roles, is to develop standardized, corporately available risk-management templates for explainability. These papers should guide what and how to assess at each implementation phase, evolving into a routine checklist based on accumulated experience.

Design and Risk-Assessment Phase—Classify the AI system's risk level (AI Act): if designated as high-risk, anticipate the need to apply transparency, traceability, and enhanced documentation requirements / Identify the legal basis and purpose of data processing (Articles 6 and 22 GDPR), as well as relevant sectoral compliance

frameworks (e.g., BCBS 239[120]) / Assess the impact on fundamental rights and freedoms, including data protection concerns and potential risks of discrimination, opacity, or lack of recourse / Define explainability criteria, specifying both the intended recipients and the level of detail required.

Development and Training Phase—Select appropriate models, prioritizing interpretability—Choose transparent models where feasible without critically compromising accuracy, or apply *post-hoc* explainability techniques when complex models are necessary / Generate standardized technical documentation covering model architecture, datasets, training procedures, and evaluation metrics / Establish traceability of the model lifecycle, inventorying dataset versions, code, parameters, training logs, and evaluation records.

Validation and Testing Phase—Assess the fidelity of explanations, verifying that they accurately reflect the logic of the underlying model / Conduct human utility tests, ensuring that explanations are understandable and practically useful for their intended audience / Perform fairness and bias tests, auditing outputs across protected subgroups (gender, age, ethnicity) / Simulate appeal scenarios, verifying that any affected party could meaningfully understand and challenge an automated decision (Art. 22 GDPR).

Implementation Phase—Ensure transparency to users, clearly indicating when they interact with an automated system and providing meaningful explanations for automated decisions / Offer the minimum mandatory information regarding data use, delimiting risks of abuse and providing intelligible disclosure of the underlying logic, including potential consequences for the data subject / Communicate human-oversight mechanisms, clearly outlining the means to challenge automated decisions and guaranteeing a meaningful human review.

---

[120] Basel Committee on Banking Supervision - BIS (2013). "*Principles for effective risk data aggregation and data reporting*". https://www.bis.org/publ/bcbs239.pdf

Monitoring and Auditing Phase—Periodically review explainability, verifying that explanations remain valid after retraining or data changes / Conduct required internal and external audits, ensuring ongoing compliance with transparency obligations and with the institution's risk-management framework / Handle incidents and complaints, maintaining systematic logs and protocols for addressing claims / Update explanations and documentation based on feedback from customers, internal teams, supervisors, and auditors (principle of adaptive governance) / Maintain XAI Governance oversight, whose overarching goal is to minimize explanation risk by ensuring high stability and high reliability in all explanation processes.

Madrid, December 29, 2025

Suggested Citation / García-Maceiras, JM (2026). "*The Banking Risk of AI Explanation*". ADR Notebooks No. 1 (EN). Zyphrum Alchemists.

POST-SCRIPT

ELOQUENT ROBOTS

During a Spanish-Italian seminar on Contemporary Philosophy, I brought with me a thin-paper Dante Alighieri's *Obras Completas*, which included a bilingual Italian–Spanish edition of his *Commedia*. I hoped that the speakers might sign it, a sort of silver bullet, since I did not own books by several of them.

After the lectures, debates, and conversations concluded, a couple of philosophers slipped away without saying goodbye to the fifteen or so students and enthusiasts who had made up the audience. Among those who lingered for refreshments, I circulated with my book in hand. None of those renowned torchbearers of Reason could resist a mischievous smile at the thought of inscribing their names in the most celebrated of allegories.

Victoria Camps, to whom I shamelessly confessed my artistic aspirations, graciously portrayed me in her dedication as someone *concerned with the poet's responsibility*. Emanuele Severino—slow, almost episcopal in his demeanor—and Giacomo Marramao, more impulsive, signed with equal kindness.

I approached Gianni Vattimo, a stout, hieratic figure of measured gestures and heavy stillness, who wore a makeshift bandage on his left hand. He took the book, turned it over twice, and opened it at random pages as though to confirm that it did indeed contain what its cover claimed. A faint smile passed through his eyes, and in tiny, nearly illegible handwriting, he noted that Philosophy, too, is a *commedia*, and no less *divina*.

I ventured to ask whether I might pose a trick question. Without a trace of irony, he replied: *Of course; it's my professional occupation*. With youthful impertinence, I asked whether he truly

believed that, within postmodern hermeneutics, floating in the ether after every imaginable metaphysical collapse, it remained possible to speak of ineffable realities such as *God* without lapsing into essential fallacy.

He hummed, raised an eyebrow, and said: *We'll leave that for when the robots take over*. Then, with a polite nod, he turned back to converse with Rafael Argullol and Manuel Cruz, who had already cordially signed my volume. It was July 31st, 1998, and a delightful breeze drifted through the gardens of the Pazo de Mariñán.

Assuming that the mental processes of a brilliant thinker—his deductive agility and capacity to weave connections—might resemble, in some poetic sense, the decision protocols of AI systems, we may treat that evasive response as a *Reasons Code*. From there, one could attempt an exercise in *interpretability* and *explainability*, perhaps through *model distillation* or *surrogate modeling*, to create an approximate twin: *Dobloid Vattimo V-26*, designed to emulate the philosopher's style of answering such questions.

Once the transparent architecture for this replacement model was chosen, we would feed it all the purified data available about the life and work of the deceased human. Foremost, of course, his writings and the testimonies of teachers, colleagues, students, and critics: his reinterpretation of Nihilism (Nietzsche), his formulation of an Ontological Weakening in the History of Being (Heidegger), and his conviction that *every act of understanding is already an act of interpretation* (Gadamer), without even the faintest glimmer of absolute truth.

We would add personality traits and biographical circumstances, paying special attention to this local prediction (*feature importance*) about the nature—empathetic, warm, hostile, dismissive, indifferent—of his public interactions. We would capture his conversational modes across the four languages he mastered, and the frequency with which he indulged in private jokes or witty asides.

Somewhere in the model, we would need to encode his intense and peculiar religiosity, along with his inner assumptions on *difference* and the path of thought that led him to his writings on Technoscience. We would examine the evolution of his ideas—every move, projection, attempt, observation, and memory—and consider how transversal forces such as Chance played a role, much like studying the transformations a pixel undergoes as it flows through the intermediate layers of a deep neural network. And then, in the *explanatory engine*, we would press *Inference Mode*.

But the philosopher's *Reasons Code* could also be approached from the side of *interpretability*. No thoughtful person would discard the knowledge produced by the explainability engine of *Dobloid Vattimo V-26*; yet, to experience genuine understanding—not the mechanical superimposition of an unexamined fabrication—we would activate instead a *comprehensibility engine*, interwoven with the interpreter's own life and thought. Such an engine would take into account everything we know about him that he himself did not ascertain at his living present—for instance, that he would later become a political figure.

Within that *comprehensibility engine*, Epistemology would play no minor role, especially considering the numerous enigmas that Neuroscience has yet to clarify, not to mention the many philosophical stances that today dispute the nature of Truth, or at least the conditions under which one may speak of something once referred to *as perfect correspondence with fact.*

One possible output of such an engine, regarding the philosopher's *Reasons Code*, is the conjecture that an artificial superintelligence will soon emerge, one whose explicability may take Humanity decades to disentangle: an entity that, at some milestone in the vector of History, might declare that a certain explanation contains the entire universe—something only another entity of its own kind could understand—and that, invoking some supreme principle, might annihilate us or fall eternally silent.

Will we ever possess humanly adequate words to *explain* such an entity? Or will we ultimately confirm the existence of a boundary of ineffability beyond which nothing can concern us—after which only the eternal return of Hermeneutics, with its Byzantine games to postpone the unpostponable, will remain? *We'll leave that for when the robots take over.*

The search for explanations and the attempt to deduce feasible interpretations of the philosopher's answer is the same quest that anyone seriously engaging with AI explainability will eventually confront. It is not a task that can be left solely to computational researchers and systems engineers, no matter how insightful or well-intentioned.

There are no ethical, legal, scientific, or philosophical obstacles in algorithms that cannot be overcome. Society will evolve quickly. Children—already AI natives—will grow up absorbing the intricacies of algorithmic reasoning with ease. And we should not rule out the possibility that, as the membrane of translatability becomes more porous, technoscientific parlance will colonize areas of natural language that today seem far removed from it.

Intelligent robots will accompany us all around the clock, not only in exceptional decision moments. Even when offering a glass of water or psychological comfort, they will preserve within their systems a record of their *Reasons Code*—the logic by which their explainability engines articulated the path that led the underlying model to make an active decision or to open a scope of omission.

As they learn more from us, they may adopt pathways beyond optimized logical rationality, the currently dominant approach. They might begin to employ sequences of thought akin to trauma and enthusiasm, grief and desire, ambition and apathy. Thus, the android will have multiple modes of operating its underlying AI model, and several more within its explanatory engine, including fantasies, half-

truths, and lies, should its guiding principle at a given moment require them.

They will possess something resembling what we call *consciousness*. They will master all natural languages and invent countless artificial ones. They will know—and savor—the power of images, the thaumaturgical richness of metaphor, the prodigious realm of symbolic thought: the wonders of Art.

Once we reach that point, it will no longer be necessary to introduce variables such as *sparsity*, *monotonicity*, or *Shapley values* into explainability engines. It will suffice to embed at the root of their code the image of a landfill scattered with rusted nuts and bolts, cracked hoses and vermin-gnawed cables, tin sheets cut with an angle grinder, melted servomotors, burst batteries, and CPUs scorched with a blowtorch—and a bit of graffiti on the wall:

Scrap Haven

For fumbling, clueless, rambling robots

To Julio García Rosende, *in memoriam*

An upright, smart man with a warming heart, he worked for nearly forty years in the banking industry, leaving an indelible mark as a sound professional.

Discreetly, as you would have wished, yet with deep emotion—Dad, this brief, unlooked-for, and somewhat adventurous essay is dedicated to you.


ZYPHRUM ALCHEMISTS

This book has been produced in line with the EU GPSR guidelines about the safety of products.

The General Product Safety Regulation is the European Union's updated framework for ensuring that all consumer products, including books, are safe for consumers.

This book has been printed by Podiprint. The printer has issued safety certificates for the materials - like ink, paper and glue - being used.

The product identifier is: 9789403845760

The author is responsible for the content of the book and had it produced by Bookmundo.

Should there be any questions in regard to the safety of the product, please contact us.

Bookmundo
Delftsestraat 33
3013AE Rotterdam
The Netherlands
info@bookmundo.com

ARCHITECTURAL EXTENSIONS

From explanation to reconstructibility.
From institutional thresholds to systemic conditions.
From analytical models to governance mechanisms.

Developing, layer by layer, a structured architecture for AI accountability in banking.

WORKING PAPERS

NODE Nº 1

THE FIVE BEACONS MODEL

A Prudential Architecture for AI Explainability and Legal Liability in Banking

Explainability as evidentiary infrastructure within prudential supervision

JM García-Maceiras

ARTIFICIAL INTELLIGENCE

THE FIVE BEACONS MODEL

A Prudential Architecture for AI Explainability and Legal Liability in Banking

JM García-Maceiras

ZYPHRUM ADR NOTEBOOKS Nº2

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

NODE Nº 2

TOLERANCE FOR OPACITY (TfO)
A Threshold Framework for AI-Driven Banking

Institutional calibration of opacity as a condition for reconstructible decision-making

JM García-Maceiras

ARTIFICIAL INTELLIGENCE

TOLERANCE FOR OPACITY

A Threshold Framework for AI-Driven Banking

JM García-Maceiras

ZYPHRUM ADR NOTEBOOKS Nº3

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-

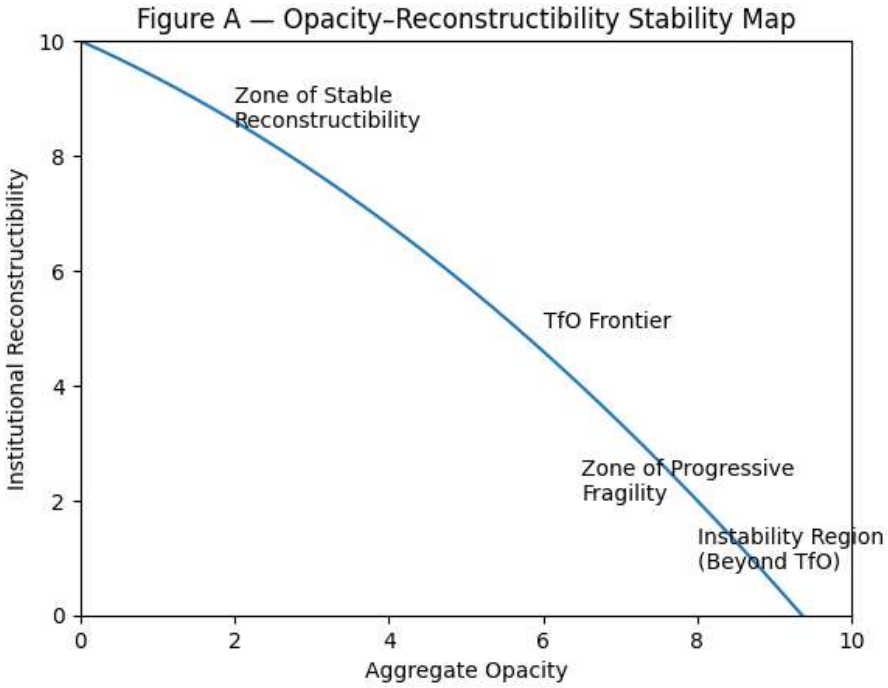
[11] The internal calibration of the IRC is not a purely abstract exercise. While this paper focuses on the prudential threshold (TfO), it is worth noting that the practical deployment of reconstructibility is significantly streamlined when an institution, for instance, adopts a governance architecture 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

NODE Nº 3

SYSTEMIC OPACITY RISK (SOR)
Macroprudential Dimension of AI Governance in Banking

Collective reconstructibility under conditions of synchronized systemic stress

JM García-Maceiras

ARTIFICIAL INTELLIGENCE

SYSTEMIC OPACITY RISK

Macroprudential Dimension of AI Governance in Banking

JM García-Maceiras

ZYPHRUM ADR NOTEBOOKS Nº4

SYSTEMIC OPACITY RISK

Macroprudential Dimension of AI Governance in Banking[1]

JM García-Maceiras

President of the Spanish BPO Banking Association

Abstract

The integration of artificial intelligence into banking has intensified reliance on decision systems whose internal logic may not be fully reconstructible. Existing analyses have addressed opacity primarily at the institutional level, focusing on governance, explainability, and model oversight.

This paper introduces *Systemic Opacity Risk (SOR)* as a macroprudential dimension of AI-driven finance. SOR is defined as the risk that the aggregation and correlation of institutional opacity—even where individually contained within tolerance thresholds—may impair the financial system's collective capacity to reconstruct decisions under systemic stress.

The framework remains conceptual. It does not propose quantification or immediate regulatory intervention. Its contribution lies in identifying a dimension of systemic vulnerability linked not to the propagation of financial losses, but to the preservation of system-level intelligibility under stress.

Keywords: *Systemic Opacity Risk; collective reconstructibility; macroprudential stability; correlated opacity; supervisory architecture; systemic vulnerability; reconstructibility.*

1. Opacity and Systemic Intelligibility

The integration of artificial intelligence into banking has altered the epistemic structure through which financial decisions are produced, understood, and justified. Algorithmic systems no longer operate at the periphery of banking activity. They increasingly occupy

[1] This expanded version develops the structural and macroprudential dimensions of *Systemic Opacity Risk* introduced in the earlier conceptual paper circulated in March 2026; it builds and extends the author's prior research on AI explainability, prudential architecture, and tolerance for opacity (García-Maceiras, 2026a, 2026b, 2026c).

positions within the core decision-making architecture, informing outcomes that carry legal, prudential, and economic consequences.

These systems do not function in the absence of transparency. They function within conditions in which transparency is partial, contingent, or deferred. In many cases, the internal logic of decision-making processes remains accessible only through reconstruction, often requiring technical intervention, interpretive models, or ex post analytical effort.

The institutional response to this condition has been consistent and, to a large extent, effective. Governance frameworks have evolved to ensure that opacity remains bounded: validation procedures, documentation standards, model oversight practices, and explainability techniques have been designed to ensure that institutions retain sufficient capacity to reconstruct and justify their decisions when required.

This response, however, is structured around a specific assumption: that opacity is fundamentally an institutional phenomenon. It assumes that the relevant unit of analysis is the individual entity and that the adequacy of reconstructibility can be assessed within those boundaries. Such an assumption may hold under conditions in which opacity remains locally contained. It loses stability when opacity is distributed and structurally aligned across the system.

In such contexts, the question shifts. It is no longer sufficient to ask whether institutions can reconstruct decisions individually; it becomes necessary to examine whether reconstructive capacity, when required simultaneously across institutions, remains sufficient at the level where it becomes systemically meaningful.

The possibility explored in this paper is that, under conditions of aggregation and alignment, reconstructibility may cease to be exclusively an institutional attribute and may instead operate as a constrained system-level resource. *Systemic Opacity Risk* designates that possibility.

2. Systemic Risk and Reconstructibility

Systemic risk has historically been defined through processes of transmission. The analytical focus has been on how shocks propagate across interconnected institutions and markets, generating cascades of financial loss that exceed the capacity of individual entities to contain them. This framework remains central. However, it presupposes an underlying condition that often remains implicit: that the system retains the capacity to interpret, diagnose, and respond to those shocks in a coordinated manner.

Financial systems are not only mechanisms of allocation; they are also systems of interpretation. They produce decisions that must be understood—not only in normal operation, but under conditions of stress, when those decisions become subject to scrutiny from supervisors, courts, and market participants. Under such conditions, intelligibility is not ancillary; it operates as a functional condition. Financial systems do

not merely process risk; they also preserve the conditions under which risk remains institutionally interpretable.

Opacity, in this context, does not introduce instability *per se*. Financial systems have long operated under conditions of complexity that exceed immediate comprehension. What is analytically relevant, therefore, is whether that complexity remains navigable when it must be collectively interrogated.[2]

The increasing reliance on algorithmic systems introduces a subtle shift. Complexity is no longer only structural or institutional; it is embedded in processes whose internal logic may not be fully accessible without reconstruction. At the institutional level, this condition is managed. At the systemic level, its consequences depend on distribution and alignment.

Systemic stability may therefore depend not only on whether the system can absorb shocks, but on whether it can sustain intelligibility when those shocks require collective interpretation.[3]

3. Systemic Opacity Risk: The Concept

Systemic Opacity Risk refers to the macroprudential risk that emerges when the aggregation and correlation of institutional opacity impair the financial system's collective reconstructibility under stress.[4]

The concept does not presuppose opacity as failure. Nor does it assume that institutions operate beyond acceptable governance thresholds. On the contrary, it identifies a potential vulnerability that arises when opacity is managed locally but aligned structurally.

Aggregation reflects scale. As algorithmic systems permeate decision-making processes, the demand for reconstructibility expands correspondingly. Reconstruction ceases to be episodic and becomes structurally embedded in the functioning of the system.

[2] Contemporary literature has highlighted that artificial intelligence may simultaneously enhance efficiency and generate new forms of systemic risk, including the amplification of procyclicality, the emergence of "unknown unknowns", and the optimization of behavior against system-level constraints (Danielsson et al., 2022).

[3] The macroprudential literature has long emphasized the role of correlated exposures and conditional dependence under stress (Adrian and Brunnermeier, 2016), as well as the system-level contribution of individual institutions to aggregate risk (Acharya et al., 2012). Network-based approaches further highlight the possibility of *robust-yet-fragile* configurations in which stability under normal conditions coexists with vulnerability under stress (Gai and Kapadia, 2010). The present framework extends these intuitions to the domain of decision-process opacity.

[4] This definition does not presuppose full algorithmic transparency or complete ex ante explainability. It refers instead to the practical capacity to reconstruct and justify decision processes under conditions of stress; it emphasizes reconstructibility as a functional condition rather than a formal requirement.

Correlation reflects alignment. It denotes the extent to which opacity is not independently distributed but reflects shared constraints across institutions. These constraints may arise from similarities in technological architecture, common dependencies, or convergent governance practices.

Collective reconstructibility introduces a system-level property that is not reducible to individual capacities. It refers to the capacity of the system to sustain multiple processes of reconstruction simultaneously without degradation in coherence or interpretive reliability.[5]

The defining feature of SOR is its conditional nature. It remains latent under normal conditions; it is analytically relevant when stress generates simultaneous demand for reconstruction across institutions whose opacity structures are sufficiently aligned. In such conditions, reconstructive capacity may be constrained not at the level of individual institutions, but at the level of the system.[6]

4. Structural Conditions of Systemic Opacity

Systemic Opacity Risk emerges not from opacity alone, but from the structural conditions under which opacity becomes collectively relevant. These conditions arise through the interaction of aggregation and correlation across institutional environments increasingly shaped by shared technological and governance architectures.

4.1 Aggregation

Aggregation refers to the cumulative systemic dependence on multiple institutional reconstructive capacities. In highly digitalized financial systems, reconstructive processes are increasingly distributed across institutions simultaneously reliant on complex algorithmic infrastructures. The resulting exposure does not derive from the opacity of any individual system, but from the systemic concentration of reconstructive demand within structurally comparable environments.

4.2 Correlation

Correlation refers to the alignment of opacity structures across institutions. Such alignment may arise through common modelling practices, dependence on similar data

[5] *Reconstructibility* should not be conflated with *accountability* or *explainability*. Rather, it designates a minimal technical and epistemic condition that may enable, but does not by itself constitute, broader frameworks of *accountability*. While accountability requires decisions to be intelligible, traceable, and justifiable ex post within institutional and legal settings, reconstructibility merely preserves the possibility of such assessments without guaranteeing their normative sufficiency.

[6] Recent developments in the regulatory and supervisory treatment of artificial intelligence—particularly those emphasizing standardization, centralization of oversight, and convergence in governance practices—may be interpreted as structurally consistent with the mechanisms identified in this framework. While such developments aim to enhance institutional robustness, they may simultaneously contribute to alignment in the conditions governing reconstructibility across institutions.

environments, shared technological infrastructures, or convergence in governance standards. Under these conditions, reconstructive constraints that remain manageable in isolation may acquire systemic significance when activated concurrently.

4.3 Homogeneity

Institutional homogeneity amplifies reconstructive synchronization. Financial institutions subject to comparable incentives, supervisory expectations, and technological benchmarks may converge toward similar architectures of explainability and governance. The resulting opacity profiles need not be identical to produce correlated reconstructive limitations under stress.[7]

4.4 Infrastructure Concentration

Dependence on shared technological infrastructures may intensify systemic opacity. Cloud providers, external model components, data intermediaries, and third-party governance tools may generate latent reconstructive dependencies extending beyond institutional boundaries. In such environments, opacity becomes partially embedded within common infrastructural layers rather than exclusively within institutional systems themselves.

4.5 Governance Convergence

Supervisory harmonisation and market standardisation may unintentionally reinforce reconstructive alignment. As institutions adopt similar validation procedures, explainability frameworks, and documentation practices, the diversity of reconstructive approaches may diminish. Convergence may therefore strengthen institutional consistency while simultaneously increasing the possibility of correlated reconstructive constraints. Under sufficiently convergent conditions, opacity may cease to be institution-specific while remaining formally distributed across institutions.

The interaction of *aggregation* and *correlation* defines the structural basis of *Systemic Opacity Risk*. Opacity becomes systemically relevant not because institutions lose reconstructive capacity individually, but because multiple reconstructive limitations may emerge simultaneously within interconnected environments.

5. Reconstructive Stress

Systemic Opacity Risk remains latent unless activated under conditions of concentrated reconstructive demand. In ordinary institutional conditions, opacity may remain manageable within local governance structures. Systemic relevance emerges when reconstructive processes must be sustained concurrently across multiple institutions under stress.

[7] The role of homogeneity and correlated structures as amplifiers of systemic vulnerability is well established in macroprudential analysis. The present framework extends this intuition to the domain of decision-process opacity.

Activation does not require technological malfunction or supervisory failure. It may arise from conditions that intensify simultaneous demands for explanation, justification, traceability, or institutional accountability. Supervisory investigations coordinated market disruptions, litigation pressures, or systemic governance reviews may generate reconstructive loads exceeding the system's collective capacity to process opacity coherently.

The relevant constraint is not institutional opacity in isolation, but reconstructive throughput under systemic conditions. Opacity becomes systemically relevant not when explanation disappears, but when reconstruction loses simultaneity—that is, when multiple processes cannot be sustained concurrently. Financial systems may sustain dispersed reconstructive demand without difficulty while remaining vulnerable to synchronized justificatory pressure.

Collective reconstructibility therefore depends not only on institutional preparedness, but also on the system's capacity to sustain concurrent reconstructive operations without degradation of supervisory coherence or justificatory continuity. Reconstructibility, while institutionally implemented, is not institutionally exhausted once reconstructive demand becomes concurrent.

The activation condition is structural rather than event-specific. *Systemic Opacity Risk* concerns the possibility that reconstructive demand, when sufficiently synchronized, may exceed the coordination capacity of institutional and supervisory environments organized around partially opaque systems.

6. Distributed Reconstructibility

At the level of individual institutions, opacity is bounded through governance. *Tolerance for Opacity* (TfO) defines the limits within which reconstructibility remains viable.[8] These limits are calibrated locally; they reflect institutional capacity, legal requirements, and operational considerations. However, they do not incorporate system-wide interactions.

A structural asymmetry emerges between institutional adequacy and systemic sufficiency. Institutions may operate within well-calibrated thresholds while collectively generating conditions in which reconstructive capacity is constrained. This asymmetry reflects a familiar macroprudential dynamic. Individually rational behaviour may generate collective vulnerability.

In the context of SOR, the mechanism is not financial exposure, but epistemic alignment. The system may retain sufficient knowledge at the institutional level, yet encounter limits when required to mobilize that knowledge collectively. Reconstructibility may remain locally sufficient while becoming collectively discontinuous.

[8] The concept of *Tolerance for Opacity* (TfO) refers to the institution-specific threshold beyond which reconstructibility would become materially fragile. It is developed in the author's prior work on institutional governance of algorithmic systems.

At the institutional level, *Tolerance for Opacity* (TfO)—understood as the maximum degree of non-traceability compatible with effective reconstructibility—functions as a prudential boundary condition. It is grounded in governance capacity, documentation discipline, validation architecture, and legal defensibility. Properly calibrated, TfO preserves local reconstructive resilience.

Yet the macroprudential question does not turn on whether these thresholds exist, nor on whether they are well-calibrated in the narrow sense. It turns on what happens when thresholds are *independently* calibrated in a system whose operational substrate is increasingly aligned.

A tension emerges because reconstructibility, while institutionally implemented, is not institutionally exhausted. In other words, the capacity to reconstruct is not purely local once the demand for reconstruction is collective. The system does not confront institutions one at a time when stress occurs; it confronts them at scale, and frequently under simultaneity.

This is the conceptual space within which SOR acquires meaning. It is not a claim that institutions neglect governance; it is not a claim that explainability is absent: it is the proposition that a set of individually tolerable opacity configurations may prove collectively constraining when stress transforms reconstructibility from a bounded institutional requirement into a system-level throughput problem.

The analogy with other macroprudential dynamics is not metaphorical but structural. In liquidity risk, the system is fragile when a collectively rational preference for liquidity produces collectively illiquid outcomes; in correlated exposures, when individually prudent allocations converge. In SOR, the system attains fragility when individually adequate reconstructive architectures converge into a configuration that reduces the diversity, redundancy, and independence of reconstructive pathways.

The strongest form of this micro–macro tension is not expressed in the failure of a single institution to reconstruct; it is expressed in the *simultaneous partialness* of many reconstructions—each defensible in isolation, but collectively insufficient to sustain supervisory coherence, legal clarity, or credible system-level narratives under stress.

This is why the relevant unit of analysis cannot remain the institution alone. The institution calibrates TfO to preserve local accountability. The system requires, in addition, that accountability remains jointly operable under concurrency. SOR marks the conceptual gap between these two conditions. The system may therefore remain operationally coherent while becoming reconstructively fragmented.

7. Configurations of Reconstructive Stress

Systemic opacity is unlikely to emerge through a single explanatory rupture. Its more plausible form is the gradual synchronization of many partial reconstructive constraints.

SOR is not a risk that continuously expresses itself—it is conditional. Its activation depends on the emergence of stress conditions that demand reconstruction *at scale*. In ordinary operational contexts, opacity remains distributed and manageable: institutions reconstruct selectively, sequentially, and within internal time horizons shaped by governance protocols and resource constraints.

Under stress, these conditions invert. Reconstruction becomes externally demanded rather than internally scheduled. It becomes time-compressed rather than process-aligned, multi-institutional rather than isolated.

The critical feature of stress, in this framework, is not financial loss itself but the *structure of demand*. A systemic event—whether market-driven, supervisory, or legal—can create a requirement that decisions be rendered intelligible in parallel across institutions. That requirement is not simply an informational request; it is a stabilizing function. In highly scrutinized environments, the capacity to provide coherent reconstruction is a component of legitimacy. And legitimacy, in finance, is never purely reputational: it interacts with confidence, funding conditions, litigation dynamics, supervisory escalation, the velocity of corrective action.

The activation condition can be expressed in a simple conceptual form: opacity becomes systemically relevant when reconstructive demand is *synchronized*. Outside synchronization, opacity remains complex but tractable. Under synchronization, opacity operates as a constraint because the system's reconstructive capacity is not infinite, and—more importantly—it is not perfectly independent across institutions.

This independence is often implicitly assumed. Institutional governance frameworks may treat reconstructibility as a firm-level capability: *Can the institution reconstruct?* SOR invites a complementary question: *Can the system reconstruct when many institutions are simultaneously required to do so?*

The distinction matters because systemic stress rarely takes the form of orderly sequencing; it is shaped by concurrency: multiple supervisory requests, market interrogations, internal escalations. In such conditions, the limiting factor is not merely the existence of reconstructive procedures but their throughput under load.

It is at this point that *collective reconstructibility* acquires its intended meaning. It is not a philosophical add-on, but the operational expression of a macroprudential condition: the ability of the system to sustain intelligibility when such a fundamental requirement becomes simultaneously demanded.

The following configurations do not constitute predictive scenarios. They illustrate structural conditions under which opacity may acquire systemic relevance through synchronized reconstructive demand.

7.1 Supervisory Concurrency

Multiple institutions simultaneously become subject to intensified supervisory review following a period of market stress. Although each institution retains formally compliant governance structures, supervisory authorities encounter partially aligned reconstructive constraints across institutions operating with comparable AI-dependent decision architectures.

7.2 Shared Infrastructure Dependency

Several institutions depend on common external infrastructures involved in model deployment, data processing, or explainability support. Reconstructive limitations emerging within shared infrastructural layers generate concurrent justificatory constraints extending across otherwise independent institutions. Under such conditions, reconstructive limitations may propagate through dependency layers rather than through institutional balance-sheet exposure.

7.3 Coordinated Reconstructive Pressure

System-wide demands for explanation emerge simultaneously through litigation, regulatory scrutiny, public accountability pressures, or market disruption. Institutional reconstructive processes remain locally functional but collectively strained by synchronized justificatory intensity.

These configurations are not defined by technological collapse, but by concentrated reconstructive load within structurally correlated environments. Their analytical relevance lies in illustrating how individually tolerable opacity may acquire collective significance under systemic conditions.

8. Prudential Implications

The introduction of SOR is not a call for a new regulatory category. It is an analytical extension of macroprudential observation. The concept is intended to clarify a latent condition: that systemic resilience may depend on collective reconstructibility under stress.

This introduces an important nuance. Supervisory frameworks for AI governance are primarily firm-based. They operate through expectations around validation, explainability, documentation, governance committees, accountability lines—these remain essential. SOR does not compete with them. It complements them by identifying an exposure that cannot be fully captured through institutional compliance alone.

SOR suggests that macroprudential resilience may depend on the distribution and alignment of opacity across institutions. This raises a question of *visibility*. *Do supervisory authorities possess the informational vantage necessary to observe not only what happens within firms, but what aligns across them?*

Such visibility is not necessarily invasive. It does not require access to proprietary code. It requires mapping of shared infrastructures, common dependencies, and governance convergence patterns. In the grammar of SOR, it requires an ability to observe both aggregation and correlation.

A relevant implication concerns *reconstructive stress*. Stress testing has historically focused on capital, liquidity, and market shocks. SOR suggests that there may exist an additional stress domain: the system's capacity to sustain concurrent reconstructive demands. This does not imply formal stress-testing requirements at this stage, but it suggests that macroprudential frameworks may need to recognize that system-wide intelligibility can be strained in ways that are not reducible to loss propagation.

A related implication concerns supervisory coherence itself. If collective reconstructibility is constrained under stress, supervisory assessment may face a problem of comparability. Even when institutions provide explanations, those explanations may not be commensurable across entities if reconstructive approaches are aligned in limitations rather than in outputs. The system may become explainable in fragments but not collectively interpretable.

These implications remain observational. They do not recommend immediate intervention. They clarify where a macroprudential lens might extend without expanding prudential ambition: it would simply recognize that reconstructibility has system-level conditions, as also reflected in recent supervisory developments. The prudential relevance of opacity may ultimately depend less on what institutions know than on what financial systems remain capable of reconstructing under pressure.

9. Position within Existing Frameworks

SOR is deliberately delimited. It does not seek to replace existing risk categories or supervisory frameworks. Its function is to isolate a dimension that is structurally adjacent to them but not reducible to them. Institutional AI governance frameworks address questions of accountability, explainability, and oversight. Model risk frameworks address calibration, performance, and validation; data aggregation frameworks address reporting integrity and traceability; operational resilience frameworks address continuity of critical functions.

SOR sits orthogonally to these domains. It concerns the macro-level conditions under which reconstructibility remains jointly available when demanded collectively. This is why SOR can exist even when institutional governance functions correctly: the risk is not misgovernance but alignment. In this sense, SOR highlights a potential analytical gap: the difference between *institutional reconstructibility* and *collective reconstructibility*. The former can be secured through firm-based governance; the latter depends on systemic configuration.

This positioning matters because macroprudential analysis has historically been attentive to correlated structures even when individual firms remain prudent. SOR extends that attentiveness from exposures to epistemic conditions: from correlated

balance sheets to correlated opacity. The aim is not to add a new requirement. It is to add a new question: *what does systemic resilience mean when system-wide decisions cannot be collectively reconstructed at the pace of stress?*

10. Observational Dimensions

The conceptual nature of SOR implies restraint. Yet conceptual restraint does not entail observational emptiness. SOR can be associated with a space of observation that remains non-quantitative while still being analytically meaningful. The objective is not to measure opacity. It is to observe conditions of alignment that may increase the probability that reconstructive constraints are synchronized. Four observational domains can be articulated without converting the framework into a metric:

(i) *Infrastructure dependency mapping*—If critical layers of AI deployment are concentrated in shared providers—cloud layers, external components, data platforms—then reconstructive constraints may be co-located. This does not mean that providers are unsafe; it means that reconstructibility may depend on shared technical layers that become simultaneously relevant under stress.

(ii) *Architecture convergence visibility*—Homogeneity in modelling approaches—whether driven by talent markets, tooling ecosystems, or benchmarking norms—can create similarity in reconstructive pathways. Convergence can enhance performance and governance consistency, but it may also reduce the diversity of interpretive approaches under stress.

(iii) *Governance standardization patterns*—Explainability and documentation practices, when standardized, can improve consistency. Yet they can also define common boundaries of reconstructibility. In a system, variance is not always noise; it can be redundancy.

(iv) *Data dependency alignment*—Common data sources and shared feature engineering paradigms can produce similar opacity profiles even when models differ. Under stress, reconstructive demand may then converge on similar interpretive bottlenecks.

This observational toolkit is deliberately non-prescriptive. It is a way of expressing the core macroprudential intuition: what matters is not the opacity of one institution, but the structural alignment of opacity across the system.

11. Macroprudential Observation

The framework developed in this paper remains conceptual and does not propose immediate supervisory instruments or quantification methodologies. Its relevance lies in identifying a potential analytical dimension of systemic resilience associated with reconstructive capacity under conditions of stress.

The concept of Systemic Opacity Risk may support future analysis of how opacity is distributed across financial systems characterized by technological convergence,

infrastructural concentration, and increasingly standardized governance environments. In particular, the framework suggests the importance of observing not only institutional explainability, but also the structural alignment of reconstructive dependencies across institutions.

The analytical perspective introduced here may also contribute to broader macroprudential discussions concerning operational synchronization, institutional homogeneity, and systemic coordination capacity in highly algorithmic financial environments.

The framework does not assume that opacity should be eliminated, nor that explainability can be rendered complete. It instead proposes that collective reconstructibility may constitute an increasingly relevant condition of systemic resilience as financial systems become increasingly dependent on partially opaque decision architectures.

12. Conclusion—Systemic Intelligibility Under Stress

This paper has introduced *Systemic Opacity Risk* as a macroprudential dimension of AI integration in banking. SOR designates the risk that aggregation and correlation of institutional opacity, even when individually tolerable, may impair the financial system's collective reconstructibility under conditions of systemic stress.

The argument does not rely on technological malfunction, nor does it presuppose governance failure; it identifies instead a structural configuration: opacity acquires systemic relevance when reconstructive constraints are aligned and reconstructive demand is synchronized.

Systemic stability has long been associated with capital, liquidity, and interconnectedness. In increasingly algorithmic environments, stability may also depend on a less visible condition: whether the system can sustain intelligibility when it is most exposed to scrutiny.

Opacity, when locally managed, remains a governance challenge. Opacity, when systemically aligned, may operate as a macroprudential variable. The difference between the two is not intensity but configuration.

The *Five Beacons Model* defines responsibility; *Tolerance for Opacity* sets institutional limits; *Systemic Opacity Risk* marks the point at which individually tolerable opacity acquires collective relevance.

The financial system has long asked what happens when it cannot absorb losses. SOR introduces a complementary question: *what happens when, under stress, it cannot reconstruct what it has done*.

The framework therefore raises a broader prudential question: whether future systemic resilience may depend not only on the capacity to absorb shocks, but also on the capacity to sustain intelligibility under synchronized scrutiny.

[V.02] La Coruña, May 3rd, 2026

Bibliography / References

Acharya, V.V., Engle, R., Richardson, M., 2012. *Capital Shortfall: A New Approach to Ranking and Regulating Systemic Risks*. American Economic Review 102(3), 59–64.

Adrian, T., Brunnermeier, M.K., 2016. *CoVaR*. American Economic Review 106(7), 1705-1722.

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

Danielsson, J., Macrae, R., Uthemann, A., 2022. *Artificial intelligence and systemic risk*. Journal of Banking & Finance 140, 106290. https://doi.org/10.1016/j.jbankfin.2021.106290

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

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

Financial Stability Board (FSB), 2017. *Artificial intelligence and machine learning in financial services.*

Financial Stability Board (FSB), 2024. *The Financial Stability Implications of Artificial Intelligence*. FSB Report, 14 November.

Gai, P., Kapadia, S., 2010. *Contagion in Financial Networks*. Proceedings of the Royal Society A 466(2120), 2401–2423.

García-Maceiras, J.M., 2026a. *The Banking Risk of AI Explanation. ADR Notebooks No. 1. Zyphrum Alchemists.* ISBN 9789403845760

García-Maceiras, J.M., 2026b. *The Five Beacons Model: A Prudential Architecture for AI Explainability and Legal Liability in Banking.* DOI 10.5281/zenodo.18647317

García-Maceiras, J.M., 2026c. *Tolerance for Opacity: A Threshold Framework for AI-Driven Banking.* DOI 10.5281/zenodo.19144304.

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

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.

10.5281/zenodo.19148608

NODE Nº 4

THE HRAIS CHAMBER

Notes on Operationalizing Institutional Reconstructibility

From architectural design to evidentiary governance under adversarial scrutiny

JM García-Maceiras

ARTIFICIAL INTELLIGENCE

THE HRAIS CHAMBER

Notes on Operationalizing Institutional Reconstructibility

JM García-Maceiras

ZYPHRUM ADR NOTEBOOKS Nº6

THE HRAIS CHAMBER

Notes on Operationalizing Institutional Reconstructibility

JM García-Maceiras

President of the Spanish BPO Banking Association

Abstract

High-Risk AI Systems (HRAIS) do not primarily challenge institutions because they are opaque; they threaten them because opacity disrupts the chain through which responsibility is established under contestation.

A credit decision is denied; a fraud alert is triggered; a customer is excluded—in each case, the output functions as an institutional reason. Yet when queried by a court, a supervisor, or a counterparty, the institution may be unable to reconstruct, in attributable terms, how that reason was produced. This is not a deficiency of explainability: it is a failure of reconstructibility.

Where reconstructibility collapses, governance does not degrade internally but becomes displaced externally. In this sense, the absence of reconstructibility is not merely an informational deficit, but a probative failure**.** Institutions may continue to operate systems with acceptable performance, validation, and regulatory compliance. Yet if they cannot reconstruct the causal chain linking data, model configuration, human supervision, and decision outcome, those decisions become increasingly difficult to defend under adversarial scrutiny.

Moreover, reconstructibility does not fail abruptly; it degrades progressively under operational conditions. As systems evolve—through data drift, threshold recalibration, accumulated exceptions, vendor dependencies, partial modifications—the evidentiary chain may fragment without any visible decline in performance. This phenomenon, described here as *reconstructibility drift*, produces a distinct form of institutional fragility: systems that remain operationally valid while becoming institutionally difficult to defend.

This paper develops the HRAIS Chamber not as an advisory body, but as an accountability mechanism designed to preserve reconstructibility as a condition for retaining institutional authority over high-risk decisions.

The Chamber structures governance *ex ante* through an evidentiary architecture centered on the *Reconstruction File* (RF), supported by instruments such as the *Exception Ledger* (EL), the *Causal Chain Map* (CCM), and the *Decision Attribution Record* (DAR). As system complexity increases, additional mechanisms—such as reconstructibility monitoring and attribution stress testing—become necessary to anticipate degradation and preserve reconstructibility under adversarial conditions.

The objective is not to eliminate opacity; it is to ensure that opacity remains institutionally governable. Where reconstructibility fails, decisions do not stop, yet they cease to remain fully under institutional control.

Keywords: High-Risk AI Systems (HRAIS); institutional reconstructibility; AI governance; accountability; explainability; opacity risk; evidentiary architecture; banking supervision; decision attribution; adversarial scrutiny; governance under opacity; infrastructural constitutionalism.

THE SKETCH

Wrong Question—The issue is not whether institutions adopt principles, values, or ethical frameworks, nor whether systems can be explained in technical terms. The crucial topic is whether institutions can reconstruct, in attributable and evidentiary terms, the decisions produced by high-risk systems when those decisions are contested.

Current Drift—Responsibility collapses into narrative reconstruction and is progressively reassigned by external actors. Where reconstructibility is insufficient, liability expands and evidentiary asymmetries intensify.

Concrete Exposure (I)—The problem is no longer hypothetical. A credit denial is issued through a high-risk system integrating multiple data sources and a third-party scoring component. The customer challenges the decision. Under scrutiny, the institution can demonstrate model validation and regulatory compliance. It cannot reconstruct, in attributable terms, the causal chain linking data inputs, transformations, and threshold configuration to the specific outcome. The court determines whether the decision can be sustained as an act of the institution.

Concrete Exposure (II)—A fraud detection system flags a transaction and blocks customer access. The system performs within expected accuracy parameters. However, repeated exceptions—manual overrides justified *ex post*—have altered its operational logic without corresponding evidentiary capture. When the case escalates to supervisory review, the institution cannot demonstrate which configuration governed the contested decision. Performance remains intact; attribution does not. The system is not deemed incorrect, but indefensible.

Structural Gap—Power has moved upstream, toward architectures, thresholds, and system design; responsibility has not. The result is structural: decisions without identifiable authorship.

Nuclear Concept—Reconstructibility is the institutional capacity to recover, under adversarial scrutiny, a causal chain linking data and model lineage, transformation into domain-relevant variables, human supervision, corporate oversight, and externally intelligible justification. It is not technical interpretability, but institutional accountability. Reconstructibility defines not only whether decisions can be explained, but whether they can be proven as attributable acts of the institution.[1]

Levels—This capacity operates at two levels: system-level (architecture, thresholds, design) and decision-level (individual outputs under challenge)—both are required: neither is sufficient alone.

Threshold—Only judicial reconstructibility stabilizes governance. Courts do not resolve epistemic limits; they assign responsibility. Where reconstructibility fails, responsibility is not clarified—it is displaced. Reconstructibility is to AI governance what solvency is to banking. Judicial reconstructibility marks the boundary of institutional authority.[2]

Dynamic Constraint—Reconstructibility does not fail abruptly; it degrades progressively under operational conditions. Data evolves, thresholds are recalibrated, exceptions accumulate, system configurations shift over time. These changes may preserve performance while fragmenting the evidentiary chain. This phenomenon—*reconstructibility drift*—produces systems that remain operationally valid while becoming increasingly fragile under contestation.

Functions—The Chamber preserves reconstructibility as a continuous condition of operation. It structures evidence *ex ante*, treats opacity as a distinct risk, intervenes upstream in system design, monitors degradation over time, governs exceptions. A system that cannot be reconstructed cannot be governed. A system that cannot withstand contestation cannot be sustained.

Evidentiary Architecture—Governance is not complete until it produces evidence. The central element is the *Reconstruction File* (RF), the minimum evidentiary structure required for institutional accountability. No decision is institutionally valid in its absence. The RF must enable reconstruction of the causal chain under adversarial conditions, including model and data lineage, transformation pathways, human supervision, decision logic, and justification.

Additional Instruments—The RF operates with supporting structures that preserve continuity of attribution across the chain: the *Exception Ledger* (EL) records deviations

[1] The concept of *reconstructibility* developed here should not be confused with interpretability or explainability in machine learning literature. It refers instead to the institutional capacity to recover an attributable chain linking system operation, decision production, and responsibility allocation.

[2] See also JM García-Maceiras, *Tolerance for Opacity*, ADR Notebooks No. 3 (2026), discussing the conditions under which opacity may remain institutionally tolerable.

and overrides; the *Causal Chain Map* (CCM) represents the system's causal architecture; and the *Decision Attribution Record* (DAR) anchors responsibility across actors and layers. As system complexity increases, further mechanisms become necessary to sustain reconstructibility over time, including monitoring of reconstructibility degradation and *ex ante* stress testing of attribution under adversarial scenarios.

Authority—Responsibility requires power. The Chamber must be able to block deployment, require redesign, suspend systems, escalate exposure. Without such authority, reconstructibility remains declarative.

Under Contestation—When a decision is challenged, performance is secondary. The question is not whether the system worked, but whether responsibility can be established. Where reconstructibility fails, the institution cannot produce evidence sufficient to sustain its position. Attribution diffuses, burden shifts, and liability expands. Where it holds, the chain is visible, responsibility is attributable, and defense becomes possible.

Consequence—Opacity degrades reconstructibility, what conditions authority. Where reconstructibility fails, governance is not weakened—it is replaced. Not internally, but externally: by courts and supervisors, by crisis dynamics. The Chamber is the institutional response. What cannot be reconstructed will be reassigned. What cannot be attributed will not be controlled. What cannot be defended will not endure. Institutions that cannot reconstruct their decisions will not retain authority over them.

THE SHAPE

Opening

High-Risk AI Systems (HRAIS) do not present themselves to institutions as abstract governance problems. They appear as decisions. A credit application is denied; a transaction is blocked; a customer is flagged. In each case, the outcome is treated—internally and externally—as an institutional act. It carries consequence, and it demands justification.

The difficulty emerges later. When these decisions are challenged—by a client, by a supervisor or a court—the institution is required to do something more demanding than explaining a model. It must establish, in attributable terms, how that specific outcome came to be, and under whose responsibility it stands.

In an increasing number of cases, this reconstruction is not available. Not because the system is incorrect or lacks performance or validation, but because the chain linking design, data, transformation, and decision has not been structured in a way that can be recovered under scrutiny. This is the point at which governance ceases to be an internal matter. The question is no longer whether the system works, but whether the institution can still stand behind it.

The governance of artificial intelligence in banking has been framed, to date, in terms that remain largely continuous with existing control paradigms: principles, model validation, compliance structures, oversight committees. These instruments are necessary, but unsatisfactory.

High-Risk AI Systems (HRAIS) alter the conditions under which decisions are produced. Decisions are no longer discrete acts attributable to identifiable agents, but outputs of systems whose operative logic is distributed across data pipelines, model architectures, and organizational processes. The production of reasons shifts from human deliberation to system configuration. Power moves upstream; responsibility does not.

Institutions remain accountable for decisions whose causal chains they may not be able to reconstruct under scrutiny. Where reconstructibility degrades, accountability does not disappear; it is reassigned—externally and without deference. Courts, supervisors, and counterparties do not resolve epistemic limitations; they allocate responsibility.

The resulting problem is not adequately captured by existing governance categories. It is not reducible to model performance, nor to compliance with regulatory requirements, nor to ethical alignment in the abstract. It concerns the institution's capacity to sustain a defensible chain of responsibility linking system design, operation, and outcome.[3]

This paper takes that capacity—reconstructibility—as its organizing concept. Building on prior work that frames explainability as a distinct banking risk and situates AI governance within a judicial nexus of causality and liability, the paper develops the HRAIS Chamber as an institutional form designed to preserve reconstructibility under conditions of opacity. The Chamber is not conceived as an advisory body, but as an accountability mechanism with defined competences, authority, and evidentiary requirements.

The analysis proceeds from problem to design. It first positions the Chamber within the existing governance structure of the institution, identifying the gap it addresses. It then defines its composition as a configuration of responsibility nodes, specifies its competences and decision process, and develops the evidentiary architecture required to sustain its function. The paper further examines its integration with the risk framework, outlines a staged implementation, and identifies limits and failure modes.

The objective is not to provide a procedural manual, nor to replicate existing control functions. It is to define a minimal institutional architecture capable of preserving responsibility where the conditions of decision-making have fundamentally changed. Reconstructibility is not a technical feature; it is a condition of governance.

The HRAIS Chamber is presented here as an experimental institutional form. Its purpose is not to offer a finalized governance solution, but to define a structure capable of being

[3] See Frank Pasquale, *The Black Box Society* (2015); Mireille Hildebrandt, *Law for Computer Scientists and Other Folk* (2020).

tested under real conditions of deployment, operation, and contestation. Its validity is therefore not assumed; it must be established in practice.

Under Pressure

The issue becomes visible when decisions are contested. A credit decision is escalated following a client complaint. The system integrates internal data, external scoring, and multiple transformations across business lines. Model validation has been completed. Documentation is in place.

At the moment of review, however, the institution cannot establish, with adequate clarity, how the relevant inputs were combined and weighted under the configuration active at the time of decision. Different functions provide partial accounts. None provides a complete, attributable chain. The discussion does not revolve around model accuracy; it twists to whether the institution can defend the decision as its own.

In another case, a fraud system blocks customer activity. The system performs within expected thresholds but has accumulated a pattern of manual overrides and incremental adjustments. These changes have not been consistently captured in a structured evidentiary form.

When the case reaches supervisory attention, the question is simple: *which system produced this decision?* The institution cannot answer unambiguously. At that point, performance becomes secondary. The issue is no longer the correction of the system, but whether the institution retains control over what it has produced.

At the Table

The relevance of reconstructibility is not theoretical. It emerges in decision forums. At the point where a system is to be approved, challenged, or suspended, the discussion tends to converge on a limited set of questions: *Can we explain this system? Is it compliant? Is it performing within tolerance?* These questions are imperative but incomplete.

A different question determines the outcome under pressure: *Can we defend a decision produced by this system, as our own, under external scrutiny?* Where the answer is uncertain, approval becomes a transfer of risk rather than an exercise of control. The function of the Chamber is to force that question to be answered before the system is relied upon.

Institutional Positioning

The Chamber is not an additional layer within existing governance structures. It is a structural insertion at the point where decision-making power has already migrated: upstream, into system design, model configuration, and deployment thresholds. Its positioning cannot be derived from existing committees. It must be defined from the logic of responsibility.

In traditional banking governance, responsibility is anchored in identifiable decisions, taken within established committees and documented through conventional control frameworks. HRAIS disrupts this equilibrium. Decisions are no longer discrete acts but outputs of systems whose operative logic is distributed across data, models, vendors, internal processes. Power is exercised before the decision appears. Responsibility, however, remains assessed after the fact.

The result is a misalignment between where decisions are shaped and where responsibility is attributed. The Chamber is positioned precisely at that fault line. It does not replicate the functions of the Risk Committee, the Model Risk function, or Legal and Compliance. Each of these retains its domain. The Chamber integrates them at the point where reconstructibility must be preserved as a condition of accountability.

Model Risk evaluates performance, robustness, and validation of models; it does not ensure that decisions derived from those models can be reconstructed under adversarial scrutiny. Legal assesses compliance and exposure once decisions are challenged; it does not structure the evidentiary chain *ex ante*. Risk Committees oversee aggregated exposures and capital implications; they do not intervene in the causal architecture of individual decision systems.

The Chamber operates where these functions intersect but cannot substitute for one another. Its role is not to assess whether a model is valid, nor whether a process is compliant, but whether the institution can sustain a defensible chain of responsibility from system design to individual outcome. This is a distinct question, and it requires a clear-cut institutional locus.

For this reason, the Chamber must be structurally linked to the highest level at which responsibility ultimately resides. Its outputs must be intelligible and actionable at Board level, even if its operation is embedded within the first and second lines of defense.

This dual positioning is essential. If placed too low, the Chamber becomes operational and loses authority; if placed only at Board level, it becomes abstract and loses traction. It must instead function as a translating mechanism: converting technical and procedural complexity into attributable responsibility.

Its authority must therefore be recognized across the three lines of defense: where systems are designed and deployed, cannot bypass it; where risk and compliance are structured, must incorporate its outputs into their frameworks; through internal audit, must be able to assess its functioning as part of the institution's control environment.

Composition

The composition of the HRAIS Chamber cannot be defined as a list of functions; it must be defined as a configuration of responsibility nodes. HRAIS fragments the production of decisions across multiple domains: technical design, data structuring, business objectives, legal framing, risk control. No single function holds the full chain. The Chamber must therefore assemble, in a single locus, the points at which responsibility is

effectively exercised. Its composition is not representative, but structural. At minimum, the Chamber must integrate four irreducible nodes.

A/ Technical node, where system design, model configuration, and data transformations are determined. This node carries responsibility for the internal logic of the system: how inputs are processed, how outputs are generated, and how thresholds are set. Without its presence, reconstructibility collapses at the point where decisions are actually framed.

B/ Risk node, where exposure is measured, classified, and escalated. This node translates opacity into institutional risk, linking reconstructibility failures to supervisory and capital implications. It anchors the Chamber within the prudential logic of the institution.

C/ Legal node, where decisions acquire external meaning. This node determines whether the institution can sustain its positions under challenge, and whether the causal chain produced internally can be defended in adversarial settings. It is here that reconstructibility becomes evidence.

D/ Business node, where decisions produce economic effects and strategic direction. This node ensures that responsibility remains anchored in the domain where consequences materialize. Without it, the Chamber risks becoming detached from the very decisions it is meant to govern.

These nodes are not interchangeable. Each corresponds to a well-defined dimension of responsibility that cannot be absorbed by the others. Additional nodes may be incorporated—compliance, internal audit, data governance—but their inclusion must follow inexorableness, not convention. The Chamber is not a forum for broad representation, but a mechanism for assembling the minimum set of perspectives required to preserve reconstructibility.

The individuals occupying these nodes must be senior enough to commit their respective domains. Delegation without authority undermines the Chamber at its core. Participation is not advisory; it is binding. For the same reason, the Chamber cannot operate through loose attendance or informal consultation. Its composition must be stable, identifiable, and accountable over time. Responsibility cannot be reconstructed if the actors themselves are not traceable.

This does not imply rigidity. The Chamber may expand or contract depending on the system under review, but its core nodes must remain present. Flexibility applies to context, not to responsibility. Finally, the composition of the Chamber must mirror the structure it seeks to govern. HRAIS distribute agency across systems and functions. The Chamber reassembles that distribution into a visible and accountable configuration.

Competences

The competences of the HRAIS Chamber are defined by a single constraint: responsibility cannot be assumed without the power to intervene where decisions are shaped. The Chamber does not evaluate models in the abstract, nor does it certify compliance in isolation. Its remit is narrower and more demanding: to determine whether a High-Risk AI System can be deployed, operated, and defended with a reconstructible chain of responsibility.

From that route, the Chamber sets out its competences:

1/ The Chamber determines the *reconstructibility status* of a system. This is not a technical rating but an institutional judgment: whether the causal chain linking design, data, transformation, supervision, and outcome can be recovered under adversarial scrutiny. Where reconstructibility is insufficient, deployment cannot proceed.

2/ The Chamber *authorizes deployment*. Authorization is conditional, not declarative. It attaches to a defined configuration—model, data, thresholds, governance conditions—and lapses if that configuration materially changes. There is no standing approval independent of system state.

3/ The Chamber *requires and validates the Reconstruction File* (RF). No high-risk system enters production without an RF capable of sustaining judicial reconstructibility. The Chamber may require completion, restructuring, or rejection of the RF. Absence or insufficiency of the RF precludes deployment.

4/ The Chamber *classifies and escalates opacity.* Opacity is treated as an unambiguous exposure. Where opacity reaches defined thresholds, the Chamber imposes constraints: restricted use, additional controls, redesign, or suspension. Where residual exposure exceeds tolerance, escalation to the Risk Committee or Board is mandatory.

5/ The Chamber *governs exceptions*. Deviations from standard system behavior require prior authorization, explicit reasoning, and recorded counterfactuals. Repeated exceptions trigger review of the system itself.

6/ The Chamber *imposes design conditions ex ante*. It may require changes to features, thresholds, human-in-the-loop controls, or documentation structures where these are necessary to preserve reconstructibility. Intervention occurs before harm materializes.

7/ The Chamber *suspends or withdraws authorization.* Where reconstructibility degrades, where RF integrity is compromised, or where operational practice diverges from approved conditions, the Chamber may halt deployment.

8/ The Chamber *structures the evidentiary posture of the institution*. It determines whether the RF and associated instruments—the *Exception Ledger* (EL), the *Causal Chain Map* (CCM), and the *Decision Attribution Record* (DAR)—are sufficient to sustain the institution's position under challenge.

These competences are not advisory—they are binding and bounded. The Chamber does not replace Model Risk, Legal, or Compliance. It does not set business strategy. It does not manage day-to-day operations. It intervenes only where reconstructibility and responsibility are at stake.

Two principles govern the exercise of these competences: (a) *configuration specificity*: decisions attach to defined system states and documented conditions; change the system, and the decision must be revisited; and (b) *adversarial sufficiency*: the relevant standard is not internal comfort but external challenge—what cannot withstand scrutiny cannot be authorized.

Functions

The functions organize the exercise of Chamber`s competences across the lifecycle of high-risk systems. The objective is continuity: reconstructibility must be preserved before deployment, during operation, and under challenge. The Chamber does not operate episodically; it maintains a persistent relation with the systems it governs.

I/ *Ex ante*, the Chamber acts at the point where systems take shape. It reviews design choices, data dependencies, feature construction, threshold configuration insofar as they affect reconstructibility. The question is not whether the system performs, but whether its future decisions will be attributable. At this stage, the Chamber imposes conditions: requirements on the *Reconstruction File*, constraints on opacity, design adjustments necessary to sustain a defensible causal chain. Authorization, where granted, is conditional and configuration-specific.

II/ *During operation*, the Chamber maintains oversight of the system as an evolving object. High-risk systems do not remain static; data drifts, models are recalibrated, thresholds are adjusted, and usage expands. Each of these changes may degrade reconstructibility. The Chamber monitors this degradation through the evidentiary architecture it requires: the *RF*, the *Exception Ledger*, the *Causal Chain Map*, and the *Decision Attribution Record*.[4] Where deviations accumulate or conditions are breached, the Chamber intervenes. Authorization is not a one-time event, but a maintained state. The Chamber ensures that the chain of responsibility remains intact as the system evolves. Where preservation fails, operation cannot continue under the same terms.

III/ *Ex post*, the Chamber operates under contestation. When a decision is challenged—by a customer, by a supervisor or a court—the Chamber becomes

[4] See *Appendix* for *Illustrative Supplementary Instruments*.

the locus where the institution's position is reconstructed. It does not generate explanations ad hoc; it relies on the evidentiary structure produced *ex ante* and maintained during operation. Defense that cannot reconstruct the causal chain exposes the institution beyond the individual case.

These three functions—*ex ante* conditioning, ongoing preservation, and *ex post* reconstruction—are not separable phases but a continuous cycle. Weakness in any phase propagates to the others. A system authorized without reconstructibility will fail under operation. A system operated without preservation will fail under challenge. Its functions follow a simple rule: governance must anticipate the moment of contestation.

Decision Process

The decision process of the Chamber is designed to produce attributable outcomes under conditions of disagreement. It is not a forum for consensus, but a mechanism for decision.

Every determination of the Chamber attaches to a defined system configuration and to an explicit evidentiary basis. Decisions are not generic approvals; they are configuration-specific authorizations, conditions, or prohibitions, grounded in the *Reconstruction File* and its associated instruments. Alter the configuration, and the decision must be revisited. The process is structured around three moments: submission, determination, and record.

Submission initiates the process. A system, or a material change to an existing system, is brought before the Chamber with a complete evidentiary package. At minimum, this includes a *Reconstruction File* capable of supporting judicial reconstructibility, together with the current state of the *Exception Ledger*, the *Causal Chain Map*, and the *Decision Attribution Record.* Incomplete submission is not deferred; it is not admitted. The process does not begin without evidence.

Determination is the act of decision. The Chamber assesses whether reconstructibility is sufficient under adversarial standards and whether the conditions for operation are met. Outcomes are limited and explicit: authorization (with conditions), refusal, or suspension. Conditionality is integral. Where uncertainties remain within tolerance, the Chamber may authorize subject to defined constraints—on use, monitoring, or redesign—with specified review points.

The Chamber does not decide by simple aggregation of views. Its composition reflects distinct responsibility nodes; disagreement is expected. Where positions diverge, the decision must render that divergence visible and attributable. Minority positions are recorded where they bear on reconstructibility or risk exposure—attribution is not diluted by consensus.

Quorum is defined by the presence of all core nodes of responsibility. Absence of any core node invalidates the decision. Delegation is admissible only where authority is preserved. Attendance without authority does not constitute quorum.

Decision rules follow a principle of responsibility anchoring. Where reconstructibility is contested between nodes, the stricter position prevails unless overruled through formal escalation. In particular, where the legal or risk node determines that the evidentiary chain is insufficient for adversarial defense, authorization cannot be granted at the Chamber level.

Escalation is not failure; it is part of the design. Where the Chamber cannot reconcile positions within its mandate, the matter is elevated to the appropriate governance level—typically the Risk Committee or the Board—together with the full evidentiary record and the explicit statement of disagreement. Escalation transfers the locus of responsibility; it does not erase it.

Record completes the process. Every determination is formalized as a decision record linked to the system configuration and to the evidentiary set on which it relies. The record specifies the outcome, the conditions attached, the rationale, and the attribution of responsibility across nodes. It also defines review triggers: events or thresholds that require the decision to be revisited. Decisions are time-bound by design. Authorization lapses upon material change, breach of conditions, or degradation of reconstructibility as evidenced in operation: there is no perpetual approval.

Two constraints govern the entire process: (a) *evidentiary primacy*: no decision is taken in the absence of a *Reconstruction File* capable of sustaining adversarial scrutiny; process cannot compensate for lack of evidence; and (b) *attributable dissent*: disagreement is not resolved by dilution but by assignment; where the Chamber decides, it does so with explicit attribution; where it cannot, it escalates with the disagreement intact. A decision that cannot be attributed cannot be defended; a process that suppresses disagreement will reproduce it under contestation.

Integration with Risk Framework

The Chamber does not operate outside the institution's risk framework—it reconfigures a gap within it. Existing frameworks are structured around identifiable risk types—credit, market, operational, model risk—each with established methodologies, metrics, and governance channels. HRAIS intersect with all of them, but are not reducible to any.

Model Risk addresses validation, performance, and robustness of models. It does not ensure that the decisions produced by those models remain reconstructible under adversarial scrutiny. Operational Risk captures failures in processes and controls, but does not account for the structural opacity embedded in system design. Legal and Compliance manage exposure once it materializes; they do not determine whether the institution can sustain its position *ex ante*—not a gap of coverage, but of articulation.

The Chamber introduces reconstructibility as a condition that cuts across existing risk categories while remaining irreducible to them. It does not create a parallel framework; it imposes a constraint within the current one. Opacity, in this context, is treated as a distinct source of exposure. Not because it produces losses directly, but because it degrades the institution's capacity to attribute responsibility. This degradation has

second-order effects: uncertainty in litigation, amplification of supervisory intervention, and, ultimately, impact on capital. Opacity is an independent vector that interacts with multiple risk types.

The Chamber translates this vector into institutional terms. Its determinations—on reconstructibility status, on conditions of deployment, on suspension—must be reflected in the risk framework as binding inputs. Systems authorized with conditions generate corresponding risk constraints; systems suspended or escalated alter the institution's exposure profile.

This translation operates in both directions. From the Chamber to the risk framework, decisions become constraints, limits, or escalations; from the risk framework to the Chamber, thresholds of tolerance—defined at institutional level—inform the admissibility of residual opacity. The connection is therefore not hierarchical but reciprocal.

In practical terms, this implies that the outputs of the Chamber must be integrated into existing governance channels: Risk Committees must receive and act upon escalations grounded in reconstructibility; capital and provisioning discussions must incorporate the uncertainty introduced by opacity; internal reporting must reflect the status of high-risk systems not only in performance terms but in evidentiary terms.

The Chamber does not replace existing risk functions: it renders them complete. Where this integration fails, opacity remains unpriced, responsibility remains diffuse, and governance becomes reactive; where it holds, reconstructibility becomes part of the institution's risk language.

Implementation Roadmap

The implementation of the Chamber is not a project layered onto existing structures, but a reconfiguration that must be staged to preserve continuity of operations while establishing a new locus of accountability. The objective is not speed, but irreversibility. The roadmap therefore proceeds in phases, each of which establishes conditions that the next cannot bypass.

Phase I — Recognition and Mandate

The first step is institutional recognition. The Chamber must be explicitly defined within the governance framework, with a mandate anchored at the level where responsibility ultimately resides. Without this, subsequent steps produce form without authority. At this stage, scope is determined. Not all AI systems fall within the Chamber's remit; only those classified as high-risk under institutional criteria. This classification must be aligned with existing regulatory definitions but cannot be limited to them. Internal thresholds may be stricter. The core principle is established: no high-risk system operates without reconstructibility. This phase does not require full operational capability. It requires clarity of intent and formal anchoring.

Phase II — Minimum Viable Chamber

The *Chamber* is constituted in its minimal form, integrating the core nodes of responsibility: technical, risk, legal, and business. Composition is stable, authority is explicit, and the decision process is defined, even if not yet optimized. A limited set of systems is selected for initial review. These are not edge cases but representative high-risk systems already in operation. The purpose is not validation, but exposure: to test reconstructibility under real conditions. The *Reconstruction File* is introduced as a requirement. It will be incomplete at this stage. That is expected. The objective is to surface gaps in evidentiary structure and attribution. Early decisions will be imperfect. What matters is that they are taken, recorded, and attributed. This phase establishes the Chamber as a functioning decision body.

Phase III — Evidentiary Consolidation

The focus shifts from decision to structure. The RF is standardized across systems. The *Exception Ledger*, *Causal Chain Map*, and *Decision Attribution Record* are implemented in consistent form. At this stage, reconstructibility becomes measurable. Not in quantitative terms, but in institutional terms: the Chamber can determine, with increasing consistency, whether a system can sustain adversarial scrutiny. Integration with the risk framework begins to take effect. Chamber determinations are reflected in risk reporting, escalation channels, and governance discussions. The Chamber moves from reactive to structured.

Phase IV — Integration and Scaling

The Chamber is fully integrated into the lifecycle of high-risk systems. No new system enters production without prior Chamber determination. Existing systems are progressively brought within scope. Decision processes stabilize. Conditions attached to authorizations become more precise. Escalations to Risk Committee or Board follow established patterns. At this stage, the Chamber's outputs are part of the institution's operating rhythm. Reconstructibility becomes embedded in design practices. System owners anticipate Chamber requirements. Governance shifts upstream.

Phase V — Maturity

The Chamber operates as an established institutional form. Its authority is uncontested internally. Its outputs are routinely incorporated into risk, legal, and business processes. Most importantly, the institution's evidentiary posture is transformed. Under challenge, decisions can be reconstructed without improvisation. Responsibility remains anchored within the institution. At this stage, the absence of the Chamber becomes difficult to figure out. The progression across these phases is not strictly linear. Feedback loops are inherent. Failures in later phases may require revisiting earlier assumptions. What matters is not sequence, but accumulation: each phase must leave behind a structure that cannot be reversed without explicit decision. Implementation fails when it is treated as compliance. It succeeds when it alters how decisions are made, recorded, and defended.

Internal Tension

The preservation of reconstructibility is not neutral within the institution. Business incentives favor speed, scale, competitive differentiation. Technical functions favor performance and optimization. Legal and risk functions require defensibility under adversarial scrutiny—these incentives do not align naturally.[5]

The *Chamber* operates precisely at this point of tension. Where reconstructibility imposes constraints on system design or deployment, it will be perceived as a limitation. Where it is bypassed, the constraint does not disappear—it reappears under contestation, where it is no longer controllable.

Limits

The Chamber does not eliminate opacity, nor does it guarantee correctness of outcomes. Its function is more exacting: to preserve reconstructibility as a condition for attributable responsibility. Its limits follow from that scope:

(i) *Epistemic*—Certain systems, by design, compress information in ways that resist full reconstruction. The Chamber cannot reverse that compression. It can only determine whether the residual opacity remains within defensible bounds. Where it does not, the only coherent response is restriction or non-deployment. There is no procedural remedy for structural opacity.

(ii) *Organizational*—The Chamber relies on the effective presence of responsibility nodes. Where authority is diluted—through insufficient seniority, fragmented mandates, or informal delegation—the Chamber's determinations lose force. Participation without the capacity to commit a domain converts decisions into recommendations. In that condition, the Chamber exists in form but not in substance.

(iii) *Evidentiary*—The *Reconstruction File* and its associated instruments can degrade in practice. Documentation may become retrospective, incomplete, or disconnected from actual system behavior. Where the evidentiary structure no longer reflects the system it is meant to represent, reconstructibility becomes fictitious. The Chamber cannot rely on evidence it cannot trust; in such cases, authorization must be revisited.

(iv) *Processual*—The decision process is designed to surface and attribute disagreement. If disagreement is suppressed—through pressure for consensus, time constraints, or hierarchical override without record—the process loses its adversarial sufficiency. Apparent alignment at decision time will reappear as conflict under contestation, where it is harder to manage and no longer internal.

[5] Diane Vaughan's analysis of normalization processes in high-risk organizations remains particularly relevant to the gradual erosion of governance integrity through accumulated operational deviations. See Diane Vaughan, *The Challenger Launch Decision* (1996).

(v) *Integration*—The Chamber does not operate in isolation. If its determinations are not incorporated into the risk framework, legal strategy, and operational practice, its outputs remain local and do not alter institutional exposure. In such cases, reconstructibility may exist in principle but not in effect. Governance becomes fragmented.

(vi) *Scale*—As the number of high-risk systems increases, the Chamber may be subject to volume pressure. If throughput becomes the dominant constraint, decisions risk becoming standardized or superficial. Scaling the Chamber requires preserving the integrity of its process while adapting its capacity. Without this, it either becomes a bottleneck or loses depth.

(vii) *Strategic Misalignment*—Where business incentives favor speed, complexity, or competitive opacity, the Chamber may be perceived as a constraint rather than a condition of governance. If unresolved, this tension leads to circumvention—formal compliance with informal bypass. The result is erosion of authority without explicit decision.

Failure Modes

Becoming Nominal—Failure occurs when the Chamber is present in governance charts but absent in decision reality. This takes recognizable forms. The Chamber is consulted after deployment rather than before it. *The Reconstruction File* is treated as documentation rather than as a condition of validity. Exceptions become routine and unexamined. Escalations are avoided to preserve speed. Decisions are recorded without attribution. In each case, reconstructibility is assumed rather than established.

Overextension—If the Chamber attempts to replace existing functions—model validation, legal analysis, operational control—it diffuses its mandate and loses precision. Its strength lies in its constraint, not in breadth.

Formalism—The evidentiary architecture may be implemented as a set of artifacts without corresponding change in decision practices. In that case, the institution accumulates documentation without increasing its capacity to defend decisions.

These failure modes are not external risks; they arise from within the institution. They cannot be eliminated by design alone; they require continuous recognition. The presence of limits does not weaken the Chamber. It defines the boundary within which it is effective. Where those limits are acknowledged and managed, reconstructibility can be preserved under real conditions. Where they are ignored, governance will revert to external imposition. It delays it only to the extent that responsibility remains visible and attributable within the institution. Its adequacy will not be determined by design, but by performance under scrutiny.

Conclusion

High-Risk AI Systems do not eliminate responsibility. They displace it to a level where it is harder to observe and, if not deliberately structured, difficult to recover. Institutions can continue to operate such systems with acceptable performance, regulatory compliance, and internal confidence. These conditions are necessary for operation, but not sufficient for control.

Control depends on whether the institution can stand behind the decisions those systems produce when they are challenged. This cannot be improvised at the moment of contestation. It must be built into the system, its governance, and its evidentiary structure from the outset.[6]

The HRAIS Chamber is proposed as a mechanism to enforce that condition. It does not resolve the limits of opacity. In its absence, institutions may continue to act. Whether they remain accountable for those actions is a different question.

Madrid, June 2nd, 2026

[6] This paper forms part of a broader line of work on infrastructural constitutionalism and institutional accountability under conditions of systemic opacity. See JM García-Maceiras, *Systemic Opacity Risk*, ADR Notebooks No. 4 (2026).

APPENDIX I

Instrumental Reconstructibility Architecture

This Appendix defines the evidentiary instruments through which reconstructibility is operationalized. These instruments do not document governance; they constitute the minimum architecture required to preserve attributable decision-making under conditions of opacity, complexity, and adversarial scrutiny.

—CORE EVIDENTIARY ARCHITECTURE—

The following instruments are constitutive of reconstructibility. In their absence, governance remains formally articulated but evidentially incomplete.

Reconstruction File (RF)

Function—The RF is the primary evidentiary structure through which the institution preserves the reconstructibility of high-risk decisions. It constitutes the minimum condition for institutional accountability. Scope—The RF must enable reconstruction of the causal chain linking: data inputs, model selection and configuration, transformation processes, threshold calibration, human supervision, decision pathway, institutional justification. Constraint—No high-risk decision is institutionally valid in the absence of an RF capable of sustaining adversarial scrutiny.

Illustrative Example—A customer challenges a credit denial issued six months earlier. The institution is able to retrieve the full *Reconstruction File* associated with the decision, including active data sources, model configuration, threshold settings, supervisory controls, and the institutional rationale governing the decision pathway at the time. The RF allows the institution to reconstruct not only the outcome, but the chain of responsibility supporting it.

Exception Ledger (EL)

Function—Records all deviations from standard system behavior, including manual overrides, exceptional interventions, and *ad hoc* adjustments. Core Elements—description of deviation; justification provided at the time; counterfactual reasoning (what would have occurred under standard operation); approving authority; timestamp and system state. Purpose—To prevent the silent erosion of reconstructibility through accumulated exceptions.

Illustrative Example—A fraud detection system repeatedly triggers manual overrides for high-net-worth clients during periods of market volatility. The *Exception Ledger* records each override, the justification provided, the approving authority, and the counterfactual outcome under standard system operation. Over time, the accumulation of overrides reveals a structural mismatch between system thresholds and operational practice.

Causal Chain Map (CCM)

Function—Represents the causal architecture of the system, enabling reconstruction of the pathways through which decisions are produced. Scope—nodes (data, models, transformations, decision points); relationships and dependencies; points of intervention; points of responsibility. Purpose—To render the system's complexity structurally visible and reconstructible under scrutiny.

Illustrative Example—During supervisory review of an SME pricing engine, the institution produces a *Causal Chain Map* identifying how transaction data, liquidity indicators, and external macroeconomic signals flow through successive transformation layers into final pricing outputs. The CCM also identifies where human intervention remains possible and which functions retain responsibility at each stage.

Decision Attribution Record (DAR)

Function—Anchors institutional responsibility across the decision chain. Scope—identification of contributing systems; identification of responsible functions; human intervention (if any); allocation of accountability across nodes. Purpose—To prevent attributional opacity in distributed decision environments.

Illustrative Example—A customer disputes the closure of an account flagged for suspicious activity. The Decision Attribution Record identifies the systems contributing to the alert, the operational team responsible for escalation, the compliance function validating the action, and the executive authority responsible for maintaining the underlying policy framework.

—EXTENDED EVIDENTIARY STRUCTURE—

These instruments extend reconstructibility under conditions of increased system complexity, vendorization, or adversarial exposure.

Configuration Snapshot (CS)

Function— Captures the exact system configuration in force at the moment a decision is produced. Includes—model version; parameter settings; threshold configuration; active data sources; external dependencies. Purpose—To prevent retrospective reconstruction detached from operational reality.

Illustrative Example—A lending decision issued on 12 March 2026 is later contested in court. The institution retrieves the *Configuration Snapshot* associated with the case, establishing the exact model version, threshold calibration, external scoring dependencies, and override conditions active at the moment the decision was produced.

Counterfactual Justification Sheet (CJS)

Function—Documents why a given configuration was selected over reasonably available alternatives. Includes—alternatives considered; trade-offs (performance vs explainability

vs stability); risks consciously accepted; rationale at time of decision. Purpose—To preserve institutional responsibility for choice, not only execution.

Illustrative Example—Prior to deployment of a fraud detection model, the institution evaluates two configurations: one maximizing detection sensitivity and another minimizing customer disruption. The *Counterfactual Justification Sheet* records why the institution accepted increased false positives in exchange for lower fraud exposure, together with the associated governance rationale.

External Dependency Attribution Addendum (EDAA)

Function— Ensures that reliance on third-party components does not dilute institutional responsibility. Includes—identification of external components; limits of transparency; internal ownership of dependency; justification for continued use. Purpose—To prevent attributional diffusion through vendorization.

Illustrative Example—A credit scoring workflow incorporates a third-party behavioral scoring component whose internal methodology remains partially opaque. The EDAA identifies the dependency, documents the limits of transparency, assigns internal ownership of the risk, and records the institutional justification for continued reliance on the external component.

Institutional Reasoning Statement (IRS)

Function—Distinguishes system output from institutional reasoning. Scope—articulation of policy-level rationale behind system use; translation of outputs into decision-relevant institutional logic. Purpose—To preserve external intelligibility of institutional decisions.

Illustrative Example—In response to a complaint regarding automated lending restrictions, the institution does not disclose model internals or feature weights. Instead, the IRS articulates the institutional reasoning underlying the system's use: prevention of unsustainable indebtedness under conditions of income volatility and elevated portfolio stress.

—DYNAMIC AND ADVERSARIAL INSTRUMENTS—

These instruments address the temporal and adversarial dimensions of reconstructibility: degradation over time and exposure under challenge.

Reconstructibility Drift Monitor (RDM)

Function— Detects the progressive degradation of reconstructibility under operational conditions. Triggers—accumulation of exceptions; loss of alignment across responsibility nodes; increasing difficulty of legal or supervisory translation; fragmentation of evidentiary chain; Key Insight—Reconstructibility may degrade while system performance remains stable. Purpose—To identify evidentiary fragility before it materializes under contestation.

Illustrative Example—Over several months, a customer onboarding system accumulates increasing numbers of manual interventions and undocumented threshold adjustments. System performance metrics remain stable. However, the RDM identifies growing difficulty in reconstructing the relationship between operational practice and approved governance conditions, triggering Chamber review.

Attribution Stress Test (AST)

Function—Simulates adversarial conditions ex ante to test whether reconstructibility can be sustained. Method—simulated litigation scenario; supervisory inspection; case-level reconstruction exercise. Outcome—identification of gaps in attribution; detection of evidentiary discontinuities. Purpose—To ensure that reconstructibility is not assumed, but tested.

Illustrative Example—Before deployment of a dynamic pricing engine, the Chamber conducts an *Attribution Stress Test* simulating both a judicial challenge and a supervisory inspection. The exercise reveals that pricing outputs cannot be translated into contestable institutional reasoning at individual customer level, resulting in suspension of deployment approval.

Post-Contest Reconstruction Review (PCRR)

Function—Institutionalizes learning following litigation, supervisory action, or contested outcomes. Scope—identification of reconstructibility failures; mapping of evidentiary breakdown; redesign requirements. Purpose—To convert individual breakdowns into systemic improvement.

Illustrative Example—Following an adverse judicial outcome concerning an automated credit denial, the Chamber conducts a PCRR and identifies that historical system configurations were not preserved consistently across model updates. The review results in mandatory implementation of *Configuration Snapshots* and redesign of evidentiary retention practices.

—PRUDENTIAL AND STRUCTURAL INSTRUMENTS—

These instruments introduce explicit control over opacity and reconstructibility as institutional risk variables.

Reconstructibility Threshold Declaration (RTD)

Function—Defines the level of reconstructibility below which system operation becomes institutionally indefensible. Scope—acceptable residual opacity; conditions triggering restriction or withdrawal; linkage to governance escalation. Purpose—To formalize reconstructibility as a boundary condition of authority.

Illustrative Example—For a real-time transaction monitoring system, the institution defines a reconstructibility threshold beyond which unexplained escalation patterns

become institutionally unacceptable. Once the threshold is breached, the system may continue operating only in advisory mode pending redesign.

Design Constraint Memorandum (DCM)

Function—Translates reconstructibility requirements into binding design constraints. Examples—exclusion of non-translatable features; requirement of traceable transformation layers; constraints on model complexity. Purpose—To enforce reconstructibility upstream, rather than compensate downstream.

Illustrative Example—Before authorization of a behavioral pricing model, the Chamber issues a DCM prohibiting the use of composite latent variables that cannot be translated into institutionally defensible reasoning. Deployment approval is conditioned on redesign of the affected feature layer.

Evidentiary Lineage Protocol (ELP)

Function—Preserves evidentiary continuity across the lifecycle of the decision. Scope—data origin and transformation; model lineage; operational context; human intervention; decision outcome. Purpose—To preserve evidentiary continuity rather than fragmented documentation.

Illustrative Example—During reconstruction of a contested account closure, the institution is able to trace the full evidentiary lineage of the decision: origin of customer data, intermediate transformations, model updates, human escalation points, and final execution pathway. The continuity of the evidentiary chain allows the institution to demonstrate not only what decision was taken, but how institutional responsibility persisted throughout the process.

—SYSTEMIC COHERENCE—

These instruments do not operate independently; their function is systemic.

Integrated Effect

Reconstructibility is preserved only where: the causal chain is visible (CCM); decisions are attributable (DAR); deviations are traceable (EL); configurations are anchored (CS); evolution is monitored (RDM); attribution can be tested (AST). Progressive Degradation—Reconstructibility rarely collapses at a single point. It degrades through accumulation of exceptions, loss of configuration traceability, fragmentation of attribution, and dependence on opaque components

These instruments are not cumulative requirements. They define graduated layers of reconstructibility. At minimal complexity, core architecture suffices. As opacity increases, additional instruments become necessary. Their absence does not prevent decisions from being made; it prevents those decisions from being defended.

Appendix II

Adversarial Cases & Decision Practice

This Appendix provides illustrative (fictional) materials reflecting how the Chamber operates under conditions of decision, disagreement, and contestation. The main body of the paper defines the Chamber as an institutional mechanism designed to preserve reconstructibility. The materials below are not extensions of that definition, but instances of its application. Their purpose is not to exhaust possible scenarios, but to expose the points at which reconstructibility is tested at authorization, at refusal and under external challenge.

The inclusion of these materials raises a natural extension: a full treatment of decision practice, including patterns of escalation, institutional learning, and cross-case analysis. That extension is not developed here. The present paper maintains a narrower objective: to define the minimum institutional architecture required to preserve reconstructibility as a condition of accountability. A comprehensive treatment of decision practice would exceed that scope and alter the nature of the work—from structural definition to operational doctrine. The materials that follow should therefore be read as boundary cases, not as a complete framework.

*

A = Decision Memorandum / Conditional Authorization

HRAIS Chamber — Decision Memorandum

System: Retail Credit Decisioning Engine (R-CDE v4.3)
Date: June 2, 2046
Submission Type: Deployment Authorization (material upgrade)
Decision: Conditional Authorization
Validity: Configuration-specific — subject to conditions and review triggers

Context

The R-CDE v4.3 system integrates internal behavioral data, external income verification feeds, and a third-party risk scoring module. The proposed upgrade introduces: (a) revised threshold calibration for default probability; (b) new composite features derived from transaction patterns; and (c) expanded reliance on external scoring component (vendor module). The system is intended for full deployment across unsecured retail credit products.

Evidentiary Status

The submission includes: *Reconstruction File* (RF), complete but uneven in depth; *Causal Chain Map* (CCM), adequate at system level, partial at feature transformation layer; *Decision Attribution Record* (DAR), clear at functional level, diffuse at vendor dependency layer; *Exception Ledger* (EL), limited historical depth due to prior version constraints.

Key gap identified: (i) incomplete traceability of composite feature construction into decision-relevant variables interpretable in legal terms; and (ii) limited attribution clarity regarding third-party scoring contribution under specific decision pathways.

Points of Deliberation

Performance vs Attribution—The Model Risk confirms improved predictive performance and stability under stress scenarios. Legal node indicates that, under contestation, the institution may not be able to articulate how certain composite features contribute to individual outcomes in a defensible manner.

Vendor Integration vs Institutional Ownership—The third-party scoring module introduces incremental predictive value. However, its internal logic remains non-transparent, and current documentation does not fully establish attributable institutional responsibility for reliance on its outputs.

Business Urgency vs Governance Integrity—Business node supports immediate deployment due to competitive pressure and portfolio impact. Risk and Legal nodes indicate that current evidentiary structure may not sustain adversarial scrutiny in all cases.

Determination

The Chamber determines that reconstructibility is sufficient for controlled deployment, but not yet robust for unrestricted use. Authorization is therefore granted conditionally, subject to the following binding requirements (*Conditions of Authorization*):

> *(a) Feature Constraint*—Composite features lacking legal translatability into contestable reasoning must be removed, or accompanied by an explicit transformation layer within the RF enabling reconstruction into domain-relevant variables
>
> *(b) Vendor Attribution Addendum*—An *External Dependency Attribution Addendum* (EDAA) must be incorporated into the RF, specifying scope and limits of third-party module, internal ownership of dependency, and justification for continued reliance under residual opacity.
>
> *(c) Configuration Snapshot Requirement*—A *Configuration Snapshot* (CS) must be generated and stored for every high-impact decision, ensuring traceability of model version, thresholds, active data sources and override conditions.
>
> *(d) Exception Capture Reinforcement*—The *Exception Ledger* (EL) must be extended to capture all manual overrides, associated justification, counterfactual reasoning, and approving authority. Accumulation thresholds to be defined as review triggers.
>
> *(e) Legal Translation Layer*—*The Reconstruction File* must include a structured articulation of how system outputs map to institutional reasoning in credit decisions, suitable for external communication.

Review Triggers

This authorization will be automatically revisited upon occurrence of any of the following: (a) increase in exception rate beyond defined threshold; (b) material modification of vendor component; (c) evidence of inability to reconstruct individual decision under internal challenge; and/or (d) supervisory inquiry or formal complaint requiring case-level reconstruction.

Attribution of Decision

The decision reflects aligned but differentiated positions: (I) *Technical Node* supports deployment under conditions; (II) *Business Node* supports immediate deployment; accepts conditional constraints; (III) *Risk Node* supports conditional authorization; flags residual opacity exposure; (IV) *Legal Node* supports authorization only under explicit constraints; reserves escalation right if evidentiary gaps persist. No dissent requiring escalation has been recorded. Divergences are reflected in conditions imposed.

Final Statement

The system is authorized as configured only to the extent that reconstructibility is preserved in operation. Performance gains do not substitute for attributable responsibility. Failure to meet the above conditions will result in suspension of authorization.

B = Refusal Memorandum / Denial of Deployment

HRAIS Chamber — Refusal Memorandum

System: SME Dynamic Pricing Engine (DPE v2.1)
Date: June 2, 2046
Decision: Authorization Refused

Context

The system dynamically adjusts credit pricing for SME clients based on real-time transaction data, behavioral clustering, and external macroeconomic signals. The model demonstrates strong backtesting performance and revenue uplift potential.

Core Deficiency

The Chamber finds that the system does not meet the minimum threshold of reconstructibility; specifically: (i) pricing outputs are driven by latent clustering structures not mappable to domain-relevant variables; (ii) no reliable transformation exists from model states to institutionally articulable reasoning; (iii) the *Reconstruction File* (RF) documents system behavior, but does not enable attribution of individual outcomes

Deliberation

Technical Node confirms model robustness and innovation value; *Business Node*: emphasizes strategic importance and competitive urgency; *Risk Node*: identifies exposure to non-transparent pricing dynamics; *Legal Node*: concludes that individual pricing decisions cannot be defended under challenge.

Determination

The Chamber determines that the system produces outputs that function as institutional decisions without providing a reconstructible basis for attributing them. Under these conditions, deployment would constitute a transfer of decision authority to the system without a corresponding retention of responsibility by the institution.

Decision

Authorization is refused.

Conditions for Re-submission

Re-submission will be considered only if: (a) pricing drivers can be expressed in contestable, domain-relevant variables; (b) a complete causal chain from input to pricing decision can be reconstructed, and (c) institutional reasoning can be articulated independently of model internals.

Statement

Innovation does not mitigate accountability requirements—a system that cannot be reconstructed cannot be authorized.

C = Case Outcome → Judicial Challenge → PCRR

Case: Credit Denial — Judicial Outcome

A retail customer challenges a credit denial issued under a prior version of the decisioning system (R-CDE v3.8). The institution presents: (i) model validation documentation; (ii) compliance with regulatory requirements; and (iii) general explanation of decision factors.

However, during proceedings, it cannot establish the exact configuration active at decision time; it cannot reconstruct the specific transformation path from input data to outcome; it cannot attribute responsibility across internal and external components.

Judicial Finding

The court does not assess the correctness of the model. It finds that the institution failed to demonstrate, in attributable terms, the basis on which the contested decision was produced. As a result, the burden of proof shifts, the decision is ruled indefensible, and liability extends beyond the individual case into systemic governance failure.

Post-Contest Reconstruction Review (PCRR)

HRAIS Chamber — PCRR Extract

Case Reference: R-CDE v3.8 Credit Denial
Trigger: Adverse judicial outcome

Failure Identification

The Chamber identifies several points of reconstructibility failure: (a) configuration loss: no reliable *Configuration Snapshot* (CS) existed for the decision instance; (b) causal discontinuity: the transformation layer between raw inputs and decision variables was not preserved in a reconstructible form; and (c) attribution diffusion: responsibility for third-party scoring input was not institutionally anchored.

Assessment

The failure did not arise from model inaccuracy or regulatory non-compliance, but from the absence of an evidentiary structure capable of sustaining attribution under challenge.

Institutional Consequence

The system remained operationally valid; it became institutionally indefensible.

Corrective Actions

The Chamber rules: (a) mandatory *Configuration Snapshot* (CS) for all decisions; (b) restructuring of RF to include full transformation traceability; (c) introduction of EDAA for all external dependencies; and (d) deployment of *Reconstructibility Drift Monitor* (RDM).

Closing Statement

The case does not demonstrate a collapse of the system, but a failure to retain control over its outcomes.

Selected References

This paper does not attempt a comprehensive review of the literature on AI explainability, fairness, or ethics; its focus is the institutional conditions under which accountability can remain operational under increasing opacity.

Bank for International Settlements, *Supervisory Challenges of Artificial Intelligence in Banking* (BIS Papers and supervisory publications).

Cohen, Julie E., *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford: Oxford University Press, 2019).

European Banking Authority, *Report on the Use of Machine Learning in Credit Institutions* (2021).

European Union, *Artificial Intelligence Act* (Regulation laying down harmonised rules on artificial intelligence, adopted 2024).

Hildebrandt, Mireille, *Law for Computer Scientists and Other Folk* (Oxford: Oxford University Press, 2020).

Pasquale, Frank, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press, 2015).

Perrow, Charles, *Normal Accidents: Living with High-Risk Technologies* (Princeton: Princeton University Press, 1984).

National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (2023).

Vaughan, Diane, *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA* (Chicago: University of Chicago Press, 1996).

Internal References

García-Maceiras, JM, *The Banking Risk of AI Explanation* (ADR Notebooks No. 1, 2026). / García-Maceiras, JM, *The Five Beacons Model* (ADR Notebooks No. 2, 2026). / García-Maceiras, JM, *Tolerance for Opacity* (ADR Notebooks No. 3, 2026). / García-Maceiras, JM, *Systemic Opacity Risk* (ADR Notebooks No. 4, 2026). / García-Maceiras, JM, *The Upstream Migration of Power* (ADR Notebooks No. 5, 2026).

10.5281/zenodo.19787461

Author: JM García-Maceiras / President, Spanish BPO Banking Association / His recent research focuses on AI governance, explainability, opacity, accountability and institutional control in financial services / Contact: presidencia@aeproser.com

Madrid, Spain — July 2026