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