

Navigating Artificial Intelligence in the Financial Sector

Practitioner's Guide to Explainability

March 12, 2025



Table of Contents

Executive Summary and Introduction	2
II. Explainability in the Financial Sector	8
III. Existing Principles of Explainability.....	15
IV. Challenges to AI Explainability	19
V. Achieving Explainability: Integrating Risk Management Practices for Trust, Reliability and Interpretability	23
Closing – other papers, prior work, moving topic	30

Executive Summary and Introduction

The adoption of generative artificial intelligence (Gen AI) in the financial sector is unlocking significant opportunities for innovation, operational efficiency, stronger resilience and enhanced customer experience. As financial institutions embrace these innovative technologies, they are also proactively addressing the complexities associated with explainability, transparency, interpretability, and trust. By leveraging their existing strengths in risk management and governance, institutions are setting a foundation for responsible and transformative Gen AI implementation.

Explainability has long been a cornerstone of model evaluation and testing in the financial industry. Traditional financial models are designed to provide clear and understandable rationales for their outputs, ensuring that stakeholders can trust the decisions made by those models and enabling ongoing testing to ensure models operate as intended. However, the advent of more sophisticated Gen AI algorithms—such as those used in fraud detection, cybersecurity, anti-money laundering and customer support—offers immense potential for enhancing efficiency and managing risks. If not approached appropriately, the lack of concrete explainability not only poses challenges to regulatory compliance but also could undermine stakeholder confidence.

As firms prepare for and implement Gen AI in their business, they are particularly focused on adhering to the principles that are the foundation for decision-making, customer communication and adherence to existing guidance and regulations. This ability to “explain” how a particular decision was arrived at, or how a service or product is delivered, requires new enhanced approaches to ensure the tenets of providing financial services are achieved.

To achieve explainability in Gen AI, financial institutions are integrating five key disciplines:

- Governance and Risk Management Frameworks (including Model Risk Management (MRM) and the NIST AI Risk Management Framework)
- Data Governance and Risk Management

- Prompting Guardrails
- Assurance and Testing
- Continuous Risk Monitoring

This paper underscores how financial institutions can fulfill the core objectives of explainability: delivering intended and trustworthy outputs, demonstrating how outcomes are derived, and ensuring transparency around their actions and results.

Regulators have consistently indicated that Model Risk Management is guidance—not regulation—and is intended to be applied flexibly based on the specific risk profile and use case of AI applications. While some financial institutions have leveraged MRM principles to manage Gen AI, regulators have also emphasized the importance of a consistent and well-structured governance approach, whether institutions apply MRM or another framework, such as the NIST AI Risk Management Framework. The NIST AI RMF, already widely referenced in the AI governance space, provides a structured yet adaptive framework for managing AI risks, aligning with regulatory expectations while allowing firms to innovate and drive AI adoption.

These guiding practices provide a clear roadmap for responsible Gen AI deployment, fostering stakeholder trust, including consumer and customer trust, addressing regulatory expectations and upholding the broader goal of maintaining accountability in AI-driven decision-making. Gen AI’s internal processing can be inherently opaque, emphasizing why explainability is crucial for preserving stakeholder confidence, meeting regulatory standards and mitigating potential risks. By tailoring explainability to each specific use case—aligning it with an institution’s risk profile, regulatory obligations and business objectives—financial services firms can integrate Gen AI into their operations responsibly. Through this approach, institutions will ensure that each implementation remains transparent, trustworthy and fully aligned with stakeholder expectations.

Importantly, this approach recognizes that by allowing for tailored strategies that align with specific business contexts and regulatory expectations, financial institutions can more effectively manage Gen AI and explain its outcomes in a manner commensurate with the associated risks.

This coordinated approach ensures that Gen AI applications align with industry standards, regulatory expectations and stakeholder trust, even when the internal processing of AI models is not fully transparent. This paper focuses specifically on Gen AI applications within financial institutions, addressing the specific unique challenges and considerations.

OBJECTIVE OF THIS APPROACH PAPER

The objective of this paper is to demonstrate that although Gen AI models currently operate as “black boxes,” financial institutions can achieve the objectives of explainability and adhere to regulatory requirements through existing risk management practices resulting in responsible implementation.

This paper explores the unique dimensions of Gen AI within the financial industry and proposes an approach that builds upon established but adaptable, mature risk management practices, effectively addressing the complexities of diverse and potentially wide-ranging Gen AI use cases.

Generative AI models come in multiple forms (Large Language Model (LLM), audio, video), such as those based on transformer architectures that understand and generate human language with an accuracy and efficiency previously unattainable, processing vast datasets to produce contextually relevant content. **Transformer architectures** are advanced deep learning models particularly effective for processing sequential data like text. They enable Gen AI models to understand context and generate coherent outputs. This type of processing, however, makes it difficult to trace outputs back to specific inputs or processes, complicating efforts to ensure meaningful explainability.

Our research and numerous meetings with AI experts in- and outside the sector identified that financial institutions could meet the core objectives that define explainability by focusing on:

- **Intended Outputs:** Clearly defining what the Gen AI model is expected to produce.
- **Trusted Results:** Ensuring outputs are reliable, accurate, and consistent through comprehensive and frequent evaluation and testing of data inputs and model outputs.
- **Interpretability from Point A to Point B:** Understanding how outputs are derived from inputs and from intended model design, real world use, and results and expectations.
- **Transparency about Processes:** Being open about the use of governance frameworks, like NIST’s AI Risk Management Framework, risk management practices, and steps to ensure compliance to enable the transparent review of relevant processes, methodologies and procedures.
- **Contextual Adaptability:** Tailoring explainability practices to each model’s intended use, specific business context and risk profile—rather than applying a uniform solution to all AI applications. This involves calibrating the depth and format of explanations based on potential impact and audience requirements (e.g., technical teams, regulators, customers).

Linking the focus areas above to the well-established and consistently implemented risk management practices below will achieve the dual purpose of advancing business goals and providing effective explanations to multiple constituents. These practices include:

- **Model Risk Management:** Assess and mitigate model risk so that Gen AI systems operate within acceptable risk parameters and align with both internal policies and regulatory expectations. The platform benefits from the expertise of Model Risk Management professionals, and business rules input is regularly challenged by Internal Audit (IA) and regulatory agencies.
 - Key Activities:

- Implement risk-based model validation, with the rigor and frequency of validation commensurate with the model's complexity and potential impact.
 - Establish clear risk thresholds and escalation procedures for AI models.
 - Conduct periodic assessments to ensure model outcomes remain accurate and fair.
 - Document assumptions, design choices, and data sources in a transparent manner for auditing and compliance.
 - **Data Governance and Risk Management:** Maintain the quality, consistency, and integrity of the data used to train, validate, and operate Gen AI models, ensuring that outputs are dependable and relevant.
 - Key Activities:
 - Implement strict data validation rules, version control, and lineage tracking to prevent data corruption or misuse.
 - Enforce privacy and security protocols to protect sensitive information.
 - Regularly review data sources to identify potential biases or inaccuracies that could propagate into Gen AI results.
 - **Prompting Guardrails:** Direct Gen AI outputs purposefully toward the intended context and desired outcomes, reducing the chance of misleading or harmful results.
 - Key Activities:
 - Provide models with carefully constructed prompts that clarify acceptable behavior, tone and outputs.
 - Embed rules or filters that detect or block risky or inappropriate content.
 - Continuously refine prompt structures based on real-world usage and feedback to maintain alignment with evolving institutional needs.
 - **Assurance and Testing:** Conduct frequent and ongoing evaluations of Gen AI models to validate accuracy, reliability, and compliance with regulatory and business requirements.
 - Key Activities:
 - Use diverse test datasets, including edge cases, to assess model performance under various scenarios.
 - Incorporate performance metrics, bias detection tools and stress-testing methodologies.
 - Perform regular audits to document compliance and demonstrate model accountability to internal and external stakeholders.
-

- **Continuous Risk Monitoring:** Detect variations from intended results and promptly mitigate emerging risks, model drift or performance degradation over time.
 - Key Activities:
 - Implement automated monitoring tools to track shifts in data distributions or model outputs.
 - Analyze real-time feedback, user complaints, and operational metrics to spot anomalies or potential biases.
 - Update or retrain models as needed, ensuring that Gen AI remains accurate, fair, and aligned with evolving regulatory expectations.

By implementing these practices and extending them to third-party risk management practices, financial services firms can enhance the transparency and trustworthiness of Gen AI applications, meeting the core objectives of explainability.

IMPORTANCE OF CONTEXT AND COMMENSURATE RISK

There is no one-size-fits-all approach to Gen AI explainability. Applying a uniform approach could lead to overly burdensome requirements for low-risk applications or insufficient rigor for high-risk ones, ultimately undermining the effectiveness and trustworthiness of Gen AI implementations. Different AI applications and use cases within the sector require varying levels of explanation and tailored risk management strategies commensurate with specific risks. Additionally, the required level and type of explanation depend significantly on the stakeholder role and purpose and the specific inquiry to which explainability is tailored.

Different stakeholders—such as regulators, internal auditors, customers and business leaders—have varying needs, understanding of the technology, and expectations regarding AI explanations. Explanations need to be appropriately detailed and relevant to the stakeholder and the specific purpose for which the explanation is sought. This contextual approach ensures that explanations are meaningful and useful, thereby facilitating trust and understanding across different audiences.

ANALOGY: HUMAN DECISION-MAKING AND GENERATIVE AI

An analogy can be drawn between human decision-making and Gen AI to illustrate this concept. When we ask a fraud analyst how they arrived at a particular decision, they do not describe the exact neural processes in their brain. Instead, they refer to their personal, business, and industry experiences; regulatory obligations and rules; the processes they followed; the data gathered and analyzed; their intended outcomes; how they tested their conclusions; and their continuous monitoring for changes in risks and outputs. This method yields trusted and interpretable results, similar to our goals with outputs from Gen AI models.

Generative AI models are inspired by the human brain. Just as we cannot pinpoint the exact neurons firing in the analyst's brain, we often cannot trace a specific decision back to a particular node in a deep learning model. Instead, the Gen AI model relies on vast amounts of training data, learned patterns and intricate internal processes, much like the analyst's accumulated knowledge and experience.

OBSTACLES TO EXPLAINABILITY ARE NOT UNIQUE TO THE FINANCIAL SYSTEM

Challenges in achieving full explainability are not unique to the use of Gen AI in the financial sector. Other industries, such as the pharmaceutical industry, face similar issues developing medications without complete knowledge of the exact mechanisms by which these medications produce therapeutic effects. Despite these uncertainties as to the exact therapeutic mechanism, pharmaceutical companies can rely on rigorous testing, clinical trials and regulatory reviews to ensure safety and efficacy when medications are made available to the public.

Similarly, underscoring the importance of robust risk management practices to ensure trust and reliability despite inherent complexities, financial institutions can employ established risk management practices to manage certain complexities of Gen AI, ensuring reliable and trustworthy outcomes even when the internal workings of the models are not fully explainable.

BALANCING RISK MANAGEMENT WITH AI INNOVATION

There is no one-size-fits-all approach to Gen AI governance. Applying an overly rigid risk management framework could unnecessarily constrain innovation, while insufficient oversight could lead to unintended risks. Financial institutions must balance risk with opportunity—tailoring AI governance to the specific business context, risk profile and use case rather than imposing a blanket approach.

A structured yet flexible governance framework is essential for ensuring accountability, transparency and adaptability in AI implementation. Institutions should adopt risk management approaches that align with their business goals, regulatory requirements and evolving AI capabilities. While Model Risk Management has been a common tool for managing AI risks, firms have multiple governance options available, including the NIST AI Risk Management Framework and other adaptable oversight structures that provide flexibility without compromising risk controls.

By emphasizing context-driven oversight, financial institutions can maximize AI-driven innovation while ensuring responsible deployment. A governance model that adapts to technological advancements and business needs will support AI adoption, maintain regulatory confidence and enhance stakeholder trust, ultimately allowing U.S. firms to accelerate innovation and drive efficiency.

II. Explainability in the Financial Sector

IMPORTANCE OF EXPLAINABILITY IN FINANCIAL SERVICES

In financial services, explainability is more than a technical requirement; it is an enabler of trust, accountability and effective decision-making. While explainability in the context of Gen AI represents new challenges, the financial sector has long emphasized related principles, such as conceptual soundness, as a cornerstone of model evaluation and validation. Regulatory frameworks have historically required clear documentation and testing to ensure models operate as intended and align with their business purposes. As institutions expand the use of Gen AI in areas like fraud detection, credit underwriting, cybersecurity and customer service, explainability builds on this legacy, ensuring that AI systems meet regulatory expectations, support informed decision-making and foster trust across stakeholders.

REGULATORY FRAMEWORKS AND GUIDELINES

This section reviews the key regulatory frameworks that have historically guided model risk management. It highlights how agencies like the OCC, Federal Reserve and FDIC have set expectations that drive not only technical rigor but also transparency and explainability in financial models.

Financial regulators such as the OCC, Federal Reserve and FDIC have long emphasized model risk management as a key guidance tool for financial oversight, while consistently clarifying that its application for AI and Gen AI should remain flexible and commensurate with risk. Their guidance—while detailed and technical—highlights the necessity for clear, explainable models that align with both internal policies and external regulatory expectations. This regulatory landscape sets the stage for our discussion on practical explainability methods, which are essential for maintaining stakeholder trust in increasingly complex Gen AI systems.

The Office of the Comptroller of the Currency's (OCC) [Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management, OCC Bulletin 2011-12](#)¹, the Federal Reserve's [SR 11-7: Guidance on Model Risk Management](#)² and the Federal Deposit Insurance Corporation (FDIC)'s [Adoption of Supervisory Guidance on Model Risk Management, Financial Institution Letter \(FIL-22-2017\)](#)³ are jointly issued supervisory guidance on Model Risk Management (MRM) and provide the same guidance, henceforth referred to as the "MRM Guidance." This guidance applies not only to models used for assessing a bank's financial status but also explicitly covers models used for stress testing and regulatory reporting.

¹ Office of the Comptroller of the Currency, "Sound Practices for Model Risk Management," OCC Bulletin 2011-12, April 4, 2011

² Federal Reserve, "SR 11-7: Guidance on Model Risk Management", April 4, 2011

³ Federal Deposit Insurance Corporation, "FIL-22-2017: Adoption of Supervisory Guidance on Model Risk Management", June 7, 2017, <https://www.fdic.gov/news/financial-institution-letters/2017/fil17022.html>

The MRM Guidance consistently underscores the importance of having the capacity to explain model outcomes. “The model methodologies and processing components that implement the theory, including the mathematical specification and the numerical techniques and approximations, should be explained in detail with particular attention to merits and limitations. Developers should ensure that the components work as intended, are appropriate for the intended business purpose, and are conceptually sound and mathematically and statistically correct. Comparison with alternative theories and approaches is a fundamental component of a sound modeling process.”⁴

Predictive AI models forecast outcomes based on historical data, similar to traditional models, focusing on accuracy, completeness and coherence. Generative AI, however, creates new content and requires additional considerations like output robustness, toxicity assessments, and performance stability. To accommodate this firms have developed specialized documentation and validation templates for Generative AI alongside our established predictive AI / Machine Learning (ML) templates.

Regarding Gen AI, explainability emphasizes the diligent review of inputs and outputs as a compensating control given the challenges in mapping inputs and outputs. Until fully explainable systems are developed, relying on these compensating controls ensures that AI models operate within acceptable parameters. This need to review inputs and outputs to ensure Gen AI’s operational effectiveness mirrors how we trust and interpret human decisions, even when the underlying neural processes are not fully understood.

MODEL RISK MANAGEMENT APPLIED TO EXPLAINABILITY

Given that model risk is a current component within the financial sector, integrating risk management practices with explainability initiatives is critical. This section explores how traditional Model Risk Management (MRM) frameworks can be adapted to the nuances of Gen AI and other complex models.

Organizations generally develop MRM frameworks aligned with the regulatory agency guidance, specifically the banking agencies’ risk management standards and the guidance for models, including the Federal Reserve and OCC’s guidance on MRM and the [Comptroller’s Handbook on Model Risk Management](#).

The OCC and Fed acknowledge that “Model risk can be diminished but not eliminated, so other tools should be used to manage model risk effectively.”⁵

The MRM Guidance offers direction to developing, validating, implementing, using and monitoring many models, including Gen AI tools.

⁴ Fed SR 11-7 & OCC 2011-12, page 6.

⁵ Office of the Comptroller of the Currency. (2011, April 4). OCC Bulletin 2011-12: Sound Practices for Model Risk Management. Retrieved <https://www.occ.gov/news-issuances/bulletins/2011/bulletin-2011-12.html#:~:text=Model%20risk%20can%20be%20diminished,with%20other%20analysis%20and%20information.>

The MRM Guidance sets out three key elements of an effective validation framework:

1. evaluation of conceptual soundness
2. ongoing monitoring
3. outcomes analysis.⁶

These elements should generally apply to many AI tools, even if their practical application may vary depending on the technology or use case.

In the MRM Guidance, “the term model refers to a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates.”⁷ The MRM Guidance notes “the definition of model also covers quantitative approaches whose inputs are partially or wholly qualitative or based on expert judgment, provided that the output is quantitative in nature.” Discussing the qualitative approaches that may not constitute models, the MRM Guidance observes that “[w]hile outside the scope of this guidance, more qualitative approaches used by banking organizations— i.e., those not defined as models according to this guidance—should also be subject to a rigorous control process.”⁸

Although the MRM Guidance may not perfectly fit Gen AI models, core principles—such as testing, context assessment and input evaluation—are broadly applicable and valuable for developing effective explainability strategies. Furthermore, regulatory expectations for validation extend to procedures even for near models—those approaches that do not strictly meet the traditional model definition. For “near models”—those that border on traditional definitions—the guidance requires that validation criteria be explicitly defined and communicated to avoid ambiguity in supervisory expectations.

COMPONENTS OF A MODEL AND SOURCES OF MODEL RISK

According to the MRM Guidance, “[a] model consists of three components: an information input component, which delivers assumptions and data to the model; a processing component, which transforms inputs into estimates; and a reporting component, which translates the estimates into useful business information.”⁹ This definition underscores the structured nature of models and the importance of managing each component effectively.

The MRM Guidance identifies two primary sources of model risk: “The use of models invariably presents model risk, which is the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports. Model risk can lead to financial loss, poor business and strategic decision making, or damage to a bank's reputation. Model risk occurs primarily for two reasons:

⁶ Fed SR 11-7 & OCC 2011-12, page 11.

- “The model may have fundamental errors and may produce inaccurate outputs when viewed against the design objective and intended business use.”¹⁰
- “The model may be used incorrectly or inappropriately. Even a fundamentally sound model producing accurate outputs consistent with the design objective of the model may exhibit high model risk if it is misapplied or misused.”¹¹

IMPORTANCE OF DATA QUALITY AND RELEVANCE

Accurate, reliable data is the lifeblood of any model. This section examines the critical role of data governance in ensuring that models are built on high-quality, relevant data, thereby enhancing both model performance and explainability.

Regarding data use within models, the MRM Guidance states: “The data and other information used to develop a model are of critical importance; there should be rigorous assessment of data quality and relevance, and appropriate documentation. Developers should be able to demonstrate that such data and information are suitable for the model and that they are consistent with the theory behind the approach and with the chosen methodology.”¹² The OCC and Fed generally do not explicitly distinguish between data obtained internally or externally from third parties in their guidance. However, the MRM Guidance emphasizes that external data—such as information from vendors or outside parties—requires particular scrutiny, especially when it pertains to new products, instruments or activities. With Gen AI models, this dynamic may evolve, as the use of diverse, large-scale datasets, often sourced externally, could introduce new considerations for assessing data quality, relevance and provenance. Financial institutions may need to adapt existing data governance practices to address these complexities effectively.

MEASURING MODEL QUALITY AND ACCOUNTING FOR UNCERTAINTY

To maintain trust in financial models, it is crucial to have measurable criteria for quality. This section outlines the key metrics and methods used to assess model accuracy, robustness and uncertainty, ensuring that models perform as intended even in the face of inherent limitations.

The MRM Guidance acknowledges that model quality is not always perfect, however model quality can be measured and assessed for specific uses. “Model quality can be measured in many ways: precision, accuracy, discriminatory power, robustness, stability, and reliability, to name a few. Models are never perfect, and the appropriate metrics of quality, and the effort that should be put into improving quality, depend on the situation.”¹³ “Accounting for model uncertainty can also include judgmental conservative adjustments to model output, placing less

¹⁰ *Id.*

¹¹ *Id.* at 4.

¹² *Id.*

¹³ *Id.* at 3.

emphasis on that model's output, or ensuring that the model is only used when supplemented by other models or approaches.”¹⁴

These types of post-hoc tools can be readily applied to assess, test, and measure results against the intended outputs, with humans in the loop to maintain oversight and control. “An integral part of model development is testing, in which the various components of a model and its overall functioning are evaluated to determine whether the model is performing as intended. Model testing includes checking the model's accuracy, demonstrating that the model is robust and stable, assessing potential limitations, and evaluating the model's behavior over a range of input values. It should also assess the impact of assumptions and identify situations where the model performs poorly or becomes unreliable.”¹⁵ If these criteria are met for Gen AI models, these compensating controls (i.e., additional measures such as independent reviews, human oversight or supplementary testing designed to mitigate residual risk) contribute to meeting the explainability objectives.

CONCEPTUAL SOUNDESS IN MODEL VALIDATION

At the core of explainability is the need for models to be conceptually sound. This section details the approaches used to validate a model's design, including documentation, testing, and comparison with alternative theories to ensure a clear and defensible model structure.

While the OCC and Fed's Supervisory Guidance on Model Risk Management do not specifically mention the term "explainability," they provide details on evaluating "conceptual soundness," which relates closely to explainability. The guidance states that conceptual soundness “involves assessing the quality of the model design and construction. It entails review of documentation and empirical evidence supporting the methods used and variables selected for the model. Documentation and testing should convey an understanding of model limitations and assumptions. Validation should ensure that judgment exercised in model design and construction is well informed, carefully considered, and consistent with published research and with sound industry practice. Developmental evidence should be reviewed before a model goes into use and as part of the ongoing validation process, in particular whenever there is a material change in the model.”¹⁶

Before deployment and throughout its lifecycle—especially after any material changes—the development process should consider producing "documented evidence in support of all model choices, including the overall theoretical construction, key assumptions, data, and specific mathematical calculations.”¹⁷ This documentation should be critically analyzed as part of the model validation, with “comparisons to alternative theories and approaches” to ensure

¹⁴ *Id.* at 7.

¹⁵ *Id.* at 6.

¹⁶ *Id.* at 11.

¹⁷ *Id.*

robustness. Special attention is given to key assumptions and variable choices, with an "analysis of their impact on model outputs and particular focus on any potential limitations."¹⁸

OCC's COMPTROLLER'S HANDBOOK ON EXPLAINABILITY

Unlike the MRM Guidance, the [OCC's Comptroller's Handbook, Model Risk Management, Version 1.0. August 2021 \("the Handbook"\)](#) emphasizes the role of explainability in the context of AI models, noting that it is a key factor in assessing the conceptual soundness of these models within a bank's model validation process. The handbook defines explainability as "the extent to which AI decisioning processes and outcomes are reasonably understood by bank personnel."¹⁹ The OCC highlights that "an evaluation of conceptual soundness may be difficult for some complex models (e.g., those that use AI approaches) because the underlying theory and logic may not be transparent."²⁰ The OCC further notes that "transparency and explainability are key considerations that are typically evaluated as part of effective risk management regarding the use of complex models."²¹

The OCC advises that "the appropriate level of explainability of a model outcome depends on the specific use and level of risk associated with that use,"²² particularly emphasizing that "models applied to significant operations or decisions (e.g., credit underwriting decisions) should be supported by thorough understanding of how the model arrived at its conclusions and validation that it is operating as intended."²³ The Handbook acknowledges the potential challenges with explaining complex models due to their complexity or "limited documentation provided for third-party models"²⁴ and suggests that "examiners should discuss with bank management the bank's process for exploring various approaches to determine whether bank personnel have an understanding of how models function and make decisions, including identifying any limitations and use of compensating controls."²⁵

IMPORTANCE OF CONTEXT AND COMMENSURATE RISK

The importance of the implementation of Gen AI in specific contexts and within particular use cases is essential when discussing AI explainability. Different applications and use cases within the financial sector present unique opportunities, risks, and challenges that must be considered when implementing and assessing Gen AI models.

As the OCC and Fed acknowledge, "[d]etails may vary from bank to bank, as practical application of this guidance should be customized to be commensurate with a bank's risk exposures, its

¹⁸ *Id.*

¹⁹ Office of the Comptroller of the Currency, *Comptroller's Handbook: Model Risk Management*, August 2021, Page 24, footnote 31.

²⁰ *Id.* at 40.

²¹ *Id.*

²² *Id.*

²³ *Id.*

²⁴ *Id.*

²⁵ *Id.*

business activities, and the complexity and extent of its model use.”²⁶ Regarding use cases and commensurate risks, MRM Guidance continues “...where models and model output have a material impact on business decisions, including decisions related to risk management and capital and liquidity planning, and where model failure would have a particularly harmful impact on a bank's financial condition, a bank's model risk management framework should be more extensive and rigorous.”²⁷ The MRM Guidance also references the use of testing applied to unique use cases, “The nature of testing and analysis will depend on the type of model and will be judged by different criteria depending on the context.”²⁸

The MRM Guidance further breaks down the contextual use based on the complexity of the materiality of models and the banking operations, “All model components, including input, processing, and reporting, should be subject to validation; this applies equally to models developed in-house and to those purchased from or developed by vendors or consultants. The rigor and sophistication of validation should be commensurate with the bank's overall use of models, the complexity and materiality of its models, and the size and complexity of the bank's operations.”²⁹

BALANCING MODEL CONTROLS: AVOIDING NEW RISKS IN GEN AI APPLICATIONS

As financial institutions increasingly adopt Gen AI, they must balance robust control measures with the need for rapid innovation. Applying a uniform set of expectations and controls across all models can inadvertently introduce new risks. For example, creditworthiness decision-making models should require more robust and granular compensating controls, which historically take considerable time to implement. If those same controls and implementation timeframes were applied to cyber and fraud tools, the considerable time to implement could expose firms and their customers to new fraud and cyber risks. For example, adversarial actors could use effective Gen AI tools to attack the institutions and their customers, while the financial sector would be hindered in deploying defensive and protective capabilities, creating an uneven playing field and increasing cyber and fraud risks to institutions and their customers.

Regulatory bodies have consistently maintained that guidance should be applied in a way that supports both effective oversight and fosters innovation. This ensures that firms are not unnecessarily constrained, allowing them to develop AI-driven solutions while maintaining sound risk management practices. Although examiner interpretations may vary, regulatory bodies strive to apply guidance consistently, which is critical to ensure that firms are not unduly constrained. It is important for the competitiveness of individual firms and the sector that supervisory practices ensure that oversight mechanisms are adaptable and that examiners are well-versed in emerging technologies.

²⁶ Fed SR 11-7 & OCC 2011-12, page 2.

²⁷ *Id.* at 5.

²⁸ *Id.* at 6.

²⁹ *Id.* at 9.

ADAPTING AI EXPLANATIONS TO MEET DIVERSE STAKEHOLDER NEEDS

Effective explainability requires tailoring information to different stakeholder groups. In addition to assessing the Gen AI model's use in a specific context, the intended audience for AI explanations, whether internal stakeholders, regulators, or customers—should influence the level of detail and type of information that needs to be provided to meet specific demands. Understanding the specific context and associated risks of each AI application is crucial for developing effective explainability practices.

- **For internal stakeholders**, detailed technical explanations are necessary to facilitate a deep understanding of the model's mechanisms and ensure robust oversight. These explanations should include information on data sources, algorithmic processes, and validation methods to enable stakeholders to evaluate the model's performance and compliance with internal policies.
- **Regulators** require clear and comprehensive documentation demonstrating how the model adheres to relevant laws and standards, including evidence of data governance practices, risk management strategies and validation results. Transparency in these areas helps build regulatory trust and confidence that AI applications meet all applicable legal requirements.
- **For customers**, the focus should be on delivering clear and relevant explanations that directly address their rights, needs, and concerns. For example, it is important for consumer-facing explanations to avoid technical jargon that creates confusion rather than clarity. The goal is to build trust by ensuring that customers can easily understand the reasoning behind AI decisions and feel confident in their ability to question or challenge those decisions if necessary.

III. Existing Principles of Explainability

Understanding the historical foundations and established definitions of AI explainability is essential for effectively implementing Gen AI in the financial sector. By examining key principles outlined by authoritative organizations like the National Institute of Standards and Technology (NIST) and the UK Information Commissioner's Office, we can align our practices with trusted standards. This section highlights the importance of these historical references and definitions, providing a framework for developing explainable AI systems that are transparent, trustworthy and compliant with regulatory expectations.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)

In the "[NISTIR 8312: Four Principles of Explainable Artificial Intelligence](#)"³⁰ publication, NIST writes: "We present four fundamental principles for explainable AI systems. These principles are heavily influenced by considering the AI system's interaction with the human recipient of the information. The requirements of the given situation, the task at hand, and the consumer will all influence the type of explanation deemed appropriate for the situation."³¹

NIST sets the stage by highlighting key concepts essential for understanding explainable AI: "we operationally define three key terms: explanation, output, and process. An explanation is the evidence, support, or reasoning related to a system's output or process. We define the output of a system as i) the outcome from or ii) the action taken by a machine or system performing a task. The output of a system differs by task...The process refers to the procedures, design, and system workflow which underlie the system."³²

The NIST document summarizes that explainable AI is a key property of trustworthy AI systems, alongside resiliency, reliability, bias and accountability. Trust in AI is influenced by how well Gen AI systems explain their decisions. Failures in explainability can lead to perceptions of bias or unfairness, slowing societal acceptance. To address this, explanations must be human-centered, understandable, accurate and reflected when the system operates outside its design.

The NIST document identifies four principles of explainable AI—Explanation, Meaningful, Accuracy, and Knowledge Limits.

NIST Fundamental Principle – Explanation

NISTIR 8312 discusses the principle of "Explanation" within the context of explainable AI, noting that that principle is satisfied if "[a] system delivers or contains accompanying evidence or reason(s) for outputs and/or processes."³³

The principles of explanation emphasize the need for AI systems to provide explanations that clarify how they reach their decisions or outputs. The primary goal is to make these explanations comprehensible and useful to the intended stakeholder who interacts with or is affected by these systems.

Key themes within the document include:

- *Comprehensibility*: Explanations should be easy to understand and should be tailored to the user's expertise and context.

³⁰ National Institute of Standards and Technology (NIST), NISTIR 8312, "Four Principles of Explainable Artificial Intelligence," available at: NISTIR 8312. Page 2.

³¹ *Id.* at 2.

- *Relevance*: Information provided should be pertinent to the user’s needs, aiding in informed decision-making.
- *Accuracy*: Explanations must accurately reflect the system's operations and reasoning to maintain trustworthiness.
- *Transparency*: The system should clearly communicate limitations or uncertainties, ensuring users are aware of potential risks or errors.

NIST Fundamental Principle – Meaningful

NISTIR 8312 discusses the principle of "Meaningful" explanations within explainable AI systems, noting that that principle is satisfied if “[a] system provides explanations that are understandable to the intended consumer(s).”³⁴ This principle emphasizes that explanations provided by AI systems should be relevant and useful to the intended audience.

The key points of this section include:

- *Audience Relevance*: Explanations should cater to the user's background, context, and needs, ensuring they are meaningful to different user groups.
- *Contextualization*: Explanations must be directly related to the user's specific application and situation.
- *Utility*: Information should empower users to make informed decisions, emphasizing practical implications and actionable insights.
- *Engagement*: Explanations should be clear, concise, and free from unnecessary complexity to enhance stakeholder trust and acceptance.

NIST Fundamental Principle - Explanation Accuracy

NISTIR 8312 discusses the principle of "Explanation Accuracy", noting that an explanation is accurate if “[a]n explanation correctly reflects the reason for generating the output and/or accurately reflects the system’s process.”³⁵ This principle emphasizes that the explanations provided by AI systems should accurately represent the processes and data used to generate the AI’s outputs.

The key aspects of this section include:

³⁴ Id.

- *Faithfulness to the Model*: Explanations must accurately reflect the AI model's operations, ensuring the explanations are not simplified or misleading.
- *Consistency with Outputs*: Explanations should directly correlate with the AI's outputs, mirroring the underlying computations and logic used in decisions or predictions.
- *Transparency of Processes*: The AI system should elucidate the data and algorithms involved, clarifying how inputs contribute to outputs to help users understand the system's reliability and potential biases.

NIST Fundamental Principle - Knowledge Limits

NISTIR 8312 addresses the principle of "Knowledge Limits" within the context of explainable AI, noting that that principle is satisfied if "[a] system only operates under conditions for which it was designed."³⁶ This principle focuses on the importance of AI systems recognizing and communicating their limitations.

Key points include:

- *Acknowledgment of Uncertainty*: AI systems should identify and communicate areas where their knowledge or data is insufficient, helping users gauge the reliability of outputs.
- *Transparent Boundaries*: AI systems should explicitly state the conditions and contexts where they perform effectively, recognizing scenarios where reliability may be compromised.
- *Error Identification*: AI systems should indicate likely error situations, especially when data is sparse or outside the training domain to help prevent misapplication of the AI system.
- *User Guidance*: Providing guidance on interpreting AI outputs, particularly in uncertain or novel situations, suggesting alternative actions or sources of information when confidence is low.
- *Alignment with User Understanding*: Explanations should balance technical accuracy with clarity and relevance, ensuring users can trust and effectively use the AI system.

THE UK INFORMATION COMMISSIONER'S SIX TYPES OF EXPLANATIONS FOR AI DECISIONS

The document “[Explaining Decisions made with AI](#)”³⁷ by the United Kingdom’s [Information Commissioner's Office \(“ICO”\)](#)³⁸, and [The Alan Turing Institute](#)³⁹, also UK-based, provide comprehensive guidelines on explaining decisions made with AI. The document emphasizes the importance of transparency, accountability and clarity in AI decision-making processes. The guidance outlines key principles and practical steps organizations can take to ensure their AI systems are understandable and justifiable to individuals affected by automated decisions.

The guidance also identifies six main types of explanations for AI decisions:

1. *Rationale Explanation*: Clarifies the reasoning behind the decision made by the AI system.
2. *Responsibility Explanation*: Specifies who is responsible for the AI decision.
3. *Data Explanation*: Details the data used to make the AI decision.
4. *Fairness Explanation*: Ensures that the decision-making process is fair and unbiased.
5. *Safety and Performance Explanation*: Assesses the reliability and performance of the AI system.
6. *Impact Explanation*: Evaluates the impact of the AI decision on the individual concerned.

IV. Challenges to AI Explainability

Despite its importance, achieving AI explainability presents several challenges, particularly in the context of Gen AI models used in the financial sector. These challenges span multiple dimensions, from the inherent complexity of the models to data quality issues and privacy concerns.

- **Model Complexity**

Sophistication of Algorithms: Gen AI models often involve highly intricate and nonlinear processes. These sophisticated algorithms can capture complex patterns in data but can be

³⁷ UK Information Commissioner’s Office, The Alan Turing Institute, NISTIR 8312, "Explaining decisions made with AI."

³⁸ The UK Information Commissioner’s Office (ICO) is the United Kingdom's regulatory body responsible for enforcing laws on data protection and freedom of information. It focuses on upholding information rights in the public interest, promoting data privacy for individuals, and ensuring transparency from public bodies. The ICO provides guidance on policy, handles complaints, and takes action against those who violate data protection laws.

³⁹ The Alan Turing Institute is the United Kingdom's national institute for data science and artificial intelligence. Established in 2015 and headquartered in the British Library, London, it brings together leading researchers and partners from academia, industry, and government to advance the field of data science and artificial intelligence, aiming to solve real-world problems through interdisciplinary research and collaboration.

difficult to interpret. These models' internal workings are often hard to access or understand, creating a “black box” that makes explainability difficult.

Lack of Transparency: The complexity of these algorithms means that even developers and data scientists may struggle to fully understand how specific outputs are generated. This lack of transparency can hinder efforts to provide clear and understandable rationales for AI decisions, which are essential for maintaining trust and regulatory compliance.

- **Data Quality Issues**

Data Dependency: The accuracy of Gen AI models heavily depends on the quality and quantity of the data on which they are trained. Poor data quality—including inaccuracies, biases or incomplete information—can obscure the decision-making process and lead to unreliable outputs. Ensuring data integrity is thus a critical component of achieving explainability.

Data Privacy: Ensuring data privacy and protecting sensitive information is paramount in the financial sector. However, these requirements further limit the extent to which Gen AI models can be made transparent and explainable. Balancing the need for explainability with stringent data privacy regulations that are continually in flux is a challenge.

- **Technical Hurdles**

Proprietary Algorithms: Many Gen AI systems use proprietary algorithms, which can hinder transparency and make it challenging to provide detailed explanations. The use of proprietary technology often means that the inner workings of these models are not disclosed, creating barriers to achieving full explainability.

Evaluation Measures: At the time of this writing, there is a lack of standardized evaluation measures for Gen AI explainability. This absence complicates the process of assessing and improving model interpretability. Developing universally accepted metrics and benchmarks for explainability remains an ongoing challenge to the deployment of Gen AI in the financial sector.

- **Ad Hoc and Non-Harmonized Regulatory Rules**

Fragmented Regulations: The patchwork of regulations applicable to Gen AI can be difficult for financial institutions to navigate. This fragmentation creates inconsistencies in compliance requirements, making it challenging for multinational banks to develop unified AI explainability strategies. The Report on Artificial Intelligence in Financial Services (U.S. Department of the Treasury)⁴⁰ notes that multiple regulatory agencies often have overlapping or diverging AI guidelines, creating a fragmented oversight environment. According to the report, this can result in inconsistent expectations surrounding model risk management and data governance, which further complicates efforts to achieve harmonized explainability practices across jurisdictions or lines of business. The Treasury also highlights that unclear or conflicting rules may hinder the effective deployment of AI models, since banks must tailor controls to satisfy differing agency priorities, ultimately slowing innovation in areas like explainability.

⁴⁰ U.S. Department of the Treasury, Artificial Intelligence in Financial Services (Washington, DC: U.S. Department of the Treasury, August 2023)

Ad Hoc Legislative Efforts: Ad hoc legislative approaches often result in rules that are not well-aligned with each other, further complicating compliance efforts and potentially stifling innovation in explainability measures. In its discussion of AI policy development, the Treasury report points out that “patchwork” or “sector-specific” regulations can lead to varying standards for AI accountability and transparency. This inconsistency forces financial institutions to make ad hoc adjustments when seeking to implement uniform explainability protocols across different product lines or geographic areas. As the report suggests, greater interagency coordination, including the development of shared definitions and best practices, could reduce duplicative or conflicting requirements. Such alignment would, in turn, encourage innovation by allowing firms to devote resources to creating robust explainability mechanisms rather than reconciling a maze of fragmented rules.

State-Level Initiatives: In the United States, individual states are increasingly enacting their own AI regulations. While these initiatives aim to address local concerns, they add another layer of complexity for banks operating across multiple states. For example, California's Consumer Privacy Act (CCPA), the Colorado AI Act (CAIA) and New York's Department of Financial Services (NYDFS) AI guidelines can impose different, and sometimes conflicting, requirements on financial institutions, particularly regarding the transparency and explainability of AI models.

Global Legislative and Regulatory Efforts: On the global stage, regulations such as the European Union's General Data Protection Regulation (GDPR) and the enacted EU AI Act seek to set high standards for AI governance. However, these laws and regulations may not align with those in other regions, leading to conflicts and additional compliance burdens for financial institutions engaged in activities beyond the U.S. A more harmonized and consistent regulatory approach would reduce these burdens and foster an environment that encourages innovation. For instance, Singapore's Monetary Authority (MAS) introduced the FEAT (Fairness, Ethics, Accountability, and Transparency) principles for AI in financial services, which can conflict with how “explainability” or “accountability” is defined under EU regulations or emerging U.S. guidelines. Even well-intentioned frameworks like MAS's can inadvertently add complexity when institutions attempt to reconcile differences in documentation requirements, model governance structures and interpretability standards across multiple jurisdictions.

Innovation Stifling: The uncertainty and variability in regulatory requirements can stifle innovation in AI explainability. Banks may be reluctant to develop and deploy new AI technologies if they are unsure about the regulatory landscape they will face. Non-harmonized regulations can lead to inconsistent enforcement practices, where similar technologies and practices might be judged differently depending on the jurisdiction.

Practical Example: With the uncertainty of how Model Risk Management (MRM) rules apply to AI models, some banks are taking a cautious approach due to the unknown of how these rules would be applied by their examiners. As a result, some banks are applying MRM Guidance to all models and, in some cases, not deploying Gen AI tools.

TECHNICAL ADVANCES IN UNDERSTANDING GENERATIVE AI MODEL OUTPUTS

Research in understanding how Gen AI models, particularly large language models (LLMs), produce outputs is evolving. Despite significant progress, it remains in its early stages. As our ability to interpret and explain AI-generated results improves, it is crucial for financial sector rules, guidance and approaches to remain adaptable. This flexibility should allow for regulatory frameworks and risk management practices to evolve in step with advancements in model explainability and transparency. Financial institutions and regulators must be prepared to adjust their strategies to align with these developments.

In Gen AI, a “feature” refers to an individual measurable property or characteristic of the data used by the model. For example, in a language model, features might include syntax, grammar, or semantics learned from training data. The model uses these features to generate responses or predictions. Efforts are underway to map these complex features more effectively. Researchers are identifying “circuits” within models, subsets of neurons and parameters responsible for specific tasks, such as sentiment analysis or syntax correction. This research aims to break down the model's operations into more understandable components, aiding in the interpretation of specific outputs.

By studying these features, researchers can gain a better understanding of how the model arrives at its outputs. For instance, a feature might be activated when the model processes a certain type of input, such as a scam email. When this feature is artificially amplified, the model might generate content aligned with that feature, demonstrating that manipulating features can change the model's behavior and better understand model outputs.

As AI technology continues to evolve, researchers must constantly update their strategies for understanding and explaining these models. This field is dynamic, and approaches to AI interpretability will likely continue to change and improve over time.

Several methods are being explored to enhance the interpretability of AI models, particularly transformer-based language models. Recent advances in explainable AI—using techniques such as Shapley Additive Explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Principled/Probabilistic Interpretability for Machine Learning (PiML)—offer practical tools to demystify complex model outputs. These methods break down AI decisions into understandable components, enabling stakeholders to more clearly see how different factors contribute to a model's result. These methods align with the two primary training paradigms of large language models (LLMs): fine-tuning and prompting. Feature attribution-based explanations are commonly used for fine-tuned LLMs to analyze how input features influence a model's outputs after training. While feature attribution can also be applied to prompted LLMs, Chain-of-Thought (CoT) prompting offers a distinct approach by eliciting reasoning steps, providing insight into how the model reaches complex decisions. Given the rapid evolution of AI, interpretability strategies must be continuously reassessed and refined to keep pace with new advancements.

V. Achieving Explainability: Integrating Risk Management Practices for Trust, Reliability and Interpretability

This section provides a set of high-level concepts that apply broadly to Gen AI in the financial sector. While these foundational practices offer a starting point for institutions seeking to adopt Gen AI responsibly and transparently, we recognize that more in-depth technical and operational details will be essential to fully implementing these strategies in practice. Future papers will expand on each of these topics, exploring more complex issues such as advanced interpretability techniques, domain-specific use cases, and emerging regulatory considerations.

A COMPREHENSIVE AND COORDINATED APPROACH

To achieve trusted, reliable and clear AI outcomes in financial services requires a coordinated approach to develop controls, matching AI use case and a firm’s assessment of risk. This approach leverages existing risk management practices, integrating them into Gen AI implementation and management processes to ensure that variation among Gen AI models meet the objectives of explainability.

An effective process involves the interplay of several core practices—each supporting the others—to design a robust approach that meets regulatory expectations and fosters stakeholder trust. Relying on a single practice, or addressing challenges, in isolation could be insufficient to reach conclusions of explainability. A coordinated practice ensures that the combination of inputs is factored into the decision making, ensuring the overall system is more resilient, trusted, and effective.

The following five risk management practices, when used together, offer trust, reliability and the clarity needed to achieve explainability:

- 1. Model Risk Management:** Strategies for Assessing and Mitigating Generative AI Model Risks
 - *Model Selection:* Set criteria for choosing foundational models to balance performance and complexity.
 - *Model Validation:* Regularly validate Gen AI models to ensure their accuracy, robustness and reliability. This process includes performance testing under various conditions to detect and correct potential issues.
 - *Documentation and Transparency:* Maintain comprehensive documentation of model development, including data sources, algorithms, and decision-making processes. For clarity, note that “compensating controls” refer to additional safeguards (e.g., manual reviews, automated validation checks or redundant risk

monitoring measures) implemented to address any residual uncertainties in the model's operation.

- *Governance Frameworks*: Establish governance frameworks that define roles and responsibilities while incorporating best practices to streamline compliance and audit functions. This may include leveraging automated workflows and integrated oversight tools to reduce administrative burdens without compromising robust risk control.
- *Expanded Scope for Model Risk*: Explicitly include stress testing, regulatory reporting, and internal limit models in risk assessments. Tailor risk management protocols to address the unique operational and regulatory challenges of these model types.
- *Explicit Validation for Near Models*: Develop specific validation criteria for models that do not strictly meet traditional definitions. Document minimum performance standards and risk thresholds to ensure that these models undergo continuous oversight.

In line with leading industry frameworks such as the NIST AI Risk Management Framework (AIRMF), many institutions are also creating model documentation packages (MDPs) that comprehensively outline the model's intended purpose, assumptions, performance metrics and stress-testing protocols.

2. Data Governance and Risk Management: Policies and Practices for Managing Data Quality and Security

- *Data Quality Management*: Implement processes to ensure the accuracy, completeness, and consistency of data used in AI models. This includes data cleaning, normalization and validation procedures.
- *Data Security*: Ensure data protection against unauthorized access and breaches through robust encryption, access controls and compliance with data protection regulations.
- *Data Lineage and Traceability*: Track the origin, movement and transformation of data throughout its life cycle to provide a clear understanding of data provenance and its impact on model outputs.
- *Data Usage*: Ensure that the data being used does not inadvertently include confidential or sensitive data and is conformance with copyright or similar limitations on use, whether from internal or external sources.

- *Data Nutritional Labeling*: Adopt a standardized format for presenting key information about AI models, similar to nutritional labels on food products. This would include details on data sources, model performance metrics, and potential biases.

To further strengthen transparency, some organizations are employing “data versioning” platforms, often part of a Machine Learning Operations (ML Ops) toolchain, that automatically track changes in training data sets to facilitate faster identification of data drift or anomalies.

3. Prompting Guardrails and Fine-Tuning Criteria: Creating guidelines to ensure AI models produce intended and fair outputs that match organizational goals and intentions.

- *Prompting Guardrails*: Establish clear prompting guardrails to ensure that AI models are used responsibly for their intended purpose. These guardrails help mitigate the risk of biased or inappropriate outputs by defining acceptable input parameters and usage contexts.
- *Fine-Tuning Criteria*: Implement fine-tuning criteria to refine Gen AI models and improve their alignment with desired outcomes. This involves frequently monitored Gen AI in production and adjusting model parameters and training data to enhance performance and reduce biases. Embedding user acceptance testing within these guardrails can help confirm that each revised prompt or fine-tuned model version meets defined performance thresholds and ethical guidelines before wider deployment.

4. Assurance and Testing: Methods for Validating AI Models and Their Outputs

- *Stakeholder Engagement*: Involve diverse stakeholders in the design and evaluation of prompting guardrails and fine-tuning processes to identify and address potential concerns and ensure that AI outputs meet organizational tolerances and regulatory requirements. Internal stakeholders may include representatives from: model risk management, fair and responsible banking, compliance, operations, risk, legal, etc. External stakeholders may include third party vendors, independent reviewers, etc.
- *Comprehensive Testing*: Conduct extensive testing of AI models under various scenarios to assess their performance, robustness, and reliability. This includes stress testing, sensitivity analysis, and scenario analysis.
- *Independent Review*: Engage independent reviewers from diverse disciplines (e.g., data science, compliance, legal and risk management) to ensure a comprehensive evaluation of AI models. This cross-discipline approach enhances objectivity,

identifies potential weaknesses and ensures alignment with technical, regulatory, and ethical standards.

5. Continuous Risk Monitoring: Techniques for Monitoring Model Drift and Evolving Risks

- *Timely Monitoring:* Implement appropriate monitoring systems to track AI model performance and detect deviations from expected behavior.
- *Ongoing Validation:* Continuously validate AI models to ensure their performance remains consistent over time and across different conditions.
- *Performance Metrics:* Use performance metrics and key indicators to identify potential model drift and other evolving risks.
- *Adaptive Management:* Develop adaptive management strategies to address emerging risks promptly and maintain model reliability and trustworthiness. For example, a bank using an AI model for credit underwriting may detect performance issues due to shifts in borrower behavior during an economic downturn. Adaptive management would involve monitoring key metrics, updating the model with recent data, and adjusting decision thresholds to ensure fair and accurate lending decisions amid changing conditions.
- *Collaborative Feedback Mechanisms:* Establish formal channels for regular feedback between regulators and banks—such as scheduled review meetings and shared best practice platforms—to continually enhance risk management strategies.

As noted above, financial institutions increasingly adopt ML Ops frameworks—such as ML flow or Kubeflow—to automate how AI models are deployed, tracked, and updated. These platforms work like a “toolbox” for managing machine learning projects, providing features such as version control for model changes, automated testing to check for errors, and alerts that signal unusual outputs. By implementing ML Ops, organizations can more easily detect anomalies or performance drift and quickly decide if a model needs re-validation or recalibration to reduce risks.

MODEL SELECTION PRACTICES: ENSURING APPROPRIATE MODELS FOR SPECIFIC USE CASES

Selecting the appropriate AI model is fundamental to achieving explainability, meeting regulatory requirements, and ensuring effective Gen AI implementation within financial institutions. Model selection practices should focus on the suitability and appropriateness of models for their specific use cases.

Key Considerations in Model Selection:

- **Alignment with Use Case Objectives:**

- *Define Clear Objectives:* Understand the specific goals of the AI application, including the desired outputs, performance metrics, and any applicable constraints.
- *Use Case Specificity:* Select models that are tailored to the particular financial domain, such as fraud detection, credit scoring, or customer service.

- **Complexity vs. Interpretability:**

- *Balancing Act:* Recognize that while complex models like deep neural networks may offer higher predictive accuracy, they often sacrifice interpretability. Striking this balance can lead to trade-offs, such as favoring simpler models in high-stakes scenarios to enhance transparency and trust, even if it means marginally lower predictive performance.
- *Prioritize Interpretability When Necessary:* For high-stakes decisions, such as credit approvals, opt for models that provide transparent and understandable outputs. These models could include well-defined structures, such as decision trees or linear regression, and incorporate human-in-the-loop interventions for oversight.

- **Regulatory Compliance:**

Meet Explainability Requirements: Ensure the chosen model can provide explanations that satisfy regulatory standards, such as under the Equal Credit Opportunity Act (ECOA) and Regulation B, which require creditors to provide clear reasons for adverse actions.

Data Considerations:

- *Data Quality and Availability:* Choose models appropriate for the volume, variety, and quality of available data. For instance, simpler algorithms (e.g., logistic regression or decision trees) may be more appropriate when data is limited or highly structured, ensuring ease of interpretation and lower risk of overfitting. In contrast, complex models like deep neural networks may be suitable when large amounts of diverse data are available, allowing for more advanced pattern recognition and predictive power.
- *Transparency in Data Usage:* Ensure that data governance practices align with the model's requirements and that data usage is transparent, ethical, and compliant with relevant guidance. For financial institutions, this can include aligning with OCC Bulletin 2011-12 / Federal Reserve SR 11-7 (Model Risk Management) for oversight of data inputs, as well as referencing the NIST AI Risk Management Framework to maintain data integrity and mitigate risks associated with AI-driven analysis.

Operational Feasibility:

- **Resource Requirements:** Consider the computational resources, expertise and time required to develop, deploy, and maintain the model.
- **Scalability and Maintenance:** Choosing models that are sustainable and scalable within the institution's operational framework.

Some financial institutions have begun to use platform-agnostic AI registries that list each deployed model's data consumption needs, monitoring cycles and resource footprints. These registries streamline the process of aligning model complexity with available infrastructure while also aiding in consistent oversight.

HUMAN OVERSIGHT IN AI SYSTEMS

Integrating human oversight into AI processes, known as “Human-in-the-Loop,” is essential for achieving explainability and building trust in Gen AI applications within the financial sector. Human experts can review, validate and interpret AI outputs, providing an additional layer of assurance that the AI system is functioning as intended and adhering to regulatory requirements.

Key Aspects:

- **Decision Validation:** Humans review AI-generated decisions and/or processes, especially in critical areas, to ensure accuracy and fairness.
- **Exception Handling:** In cases where the AI model is uncertain or encounters novel situations, human judgment is used to make the final decision.
- **Continuous Improvement:** Feedback from human reviewers is used to refine and improve AI models over time.
- **Examiner Training for Innovation:** Implement ongoing training programs for examiners on emerging Gen AI technologies, ensuring that oversight is both consistent and innovation-friendly.

Practical Examples:

Below are use cases demonstrating how Human-in-the-Loop oversight can strengthen AI applications in banking and finance:

- **Fraud Detection:** AI models flag potentially fraudulent transactions—e.g., unusual credit card charges or suspicious digital payments. Human fraud analysts review the flagged transactions, apply their domain expertise (e.g., customer transaction histories, known fraud patterns) and make the final determination on whether to escalate or clear the alerts.

- *AML/Transaction Monitoring*: Models identify anomalies in transaction patterns that might indicate money laundering or other illicit activities. Human compliance teams investigate the higher-risk alerts, verifying that the flagged behaviors either warrant regulatory reporting (e.g., Suspicious Activity Reports) or can be safely dismissed.
- *Payment Verification*: When AI systems detect irregularities in payment flows—such as mismatched account information or unusually high transfer amounts—human payment specialists verify authenticity and ensure compliance with internal controls before transactions are finalized.
- *Credit Decisioning*: While AI models evaluate loan and credit card applications by analyzing credit scores, transaction histories and other data, human underwriters review borderline or high-stakes decisions. This ensures regulatory requirements (e.g., adverse action notices) are met and that unique borrower circumstances are considered.
- *Trading & Liquidity Management*: AI tools generate trade recommendations or liquidity forecasts based on real-time market data. Human traders and treasury managers review these suggestions, apply market intuition and strategic considerations, and decide whether to execute or adjust AI-driven strategies.
- *Forecasting & Business Modeling*: AI systems predict future revenue streams, market trends, and budget requirements. Human financial planners and senior managers validate these models' assumptions, ensuring that strategic decisions incorporate business context and financial and risk management principles.
- *Marketing & Customer Support*: AI-driven customer segmentation and message personalization can improve marketing campaigns. However, human marketing specialists oversee content accuracy and relevance, ensuring compliance with privacy regulations and alignment with brand guidelines.
- *Cybersecurity*: AI systems detect anomalies in network traffic and user behavior, signaling potential cyber threats. Security specialists then investigate these flagged events to confirm threats, trigger appropriate incident response actions and continuously improve detection rules.
- *Analytics*: AI-powered platforms aggregate and analyze large volumes of financial news, market data, and internal performance metrics. Human analysts assess the insights provided, verifying relevance and accuracy before they inform risk assessments or strategic decisions.

The Human-in-the-Loop approach aligns with regulatory expectations around explainability by ensuring that human judgment and expertise are integrated into the AI decision-making process. This involvement helps to mitigate risks associated with AI models' limitations and potential biases, enhancing the overall transparency and accountability of the system. By incorporating human oversight, financial institutions demonstrate their commitment to responsible AI deployment and to maintaining the necessary level of explainability required by regulators.

Human-in-the-Loop approaches enhance explainability by:

- *Providing clear rationales for decisions:* Human reviewers can refine and customize the explanations of AI-driven outcomes, ensuring that each stakeholder—whether a regulator, customer or internal auditor—receives the appropriate level of detail. This tailored communication helps clarify why a particular recommendation or decision was made, bolstering trust, and understanding.
- *Ensuring that ethical considerations are included in the decision-making process:* By incorporating human judgment, organizations can uniquely calibrate AI-driven decisions to align with their specific values, risk tolerances and domain requirements. This human oversight allows for context-sensitive ethical evaluations, preventing one-size-fits-all outcomes and supporting more responsible deployments.
- *Allowing for accountability:* Assigning a human reviewer as the final decision-maker creates a clear chain of responsibility. In the event of errors, disputes or regulatory inquiries, it is evident who approved or overrode the AI’s recommendation, thereby reinforcing transparency, governance, and adherence to compliance obligations.

Closing – other papers, prior work, moving topic

As the financial industry embraces generative AI, meeting the objectives of explainability remains paramount to ensure both responsible and effective implementation. The rapidly evolving nature of AI technologies and the emergence of new use cases call for continual collaboration across the industry. By sharing insights, approaches, and strategies, financial institutions can collectively address the complexities of AI explainability and adapt on a firm-to-firm basis to the technological advancements in reshaping financial services. This collaborative effort is reflected in initiatives like the Bank Policy Institute's "Navigating Artificial Intelligence in Banking"⁴¹ and our response to the Treasury's Request for Information on Uses, Opportunities, and Risks of Artificial Intelligence in the Financial Services Sector.⁴²

These activities underscore not only the importance of adaptability and continuous improvement in AI governance and risk management practices, but also the historic and ongoing engagement between the industry and regulators. This engagement helps protect the broader financial sector and the communities it serves by ensuring that evolving best practices continue to align with regulatory expectations and emerging technologies.

Ongoing Collaboration: Regulators and financial institutions should establish regular feedback sessions and joint working groups to continuously refine model risk management practices. Periodic updates to this guidance, informed by industry input and emerging technological trends, will help ensure that supervisory practices evolve alongside innovation.

⁴¹ Bank Policy Institute. (April 2024). Navigating Artificial Intelligence in Banking. Retrieved from <https://bpi.com/wp-content/uploads/2024/04/Navigating-Artificial-Intelligence-in-Banking.pdf>.

⁴² Bank Policy Institute. (August 2024). "Response to Treasury's AI Request for Information." Retrieved from <https://bpi.com/wp-content/uploads/2024/08/BPI-Treasury-AI-RFI-Response-2024-4878-9975-4705-v10.pdf>.

As outlined in this paper, leveraging established practices, and tailoring them to generative AI, can deliver transparent, reliable and auditable outcomes—without requiring every aspect of the model to be fully “white box.” By maintaining a firm commitment to clarity, human oversight, and contextual understanding, the financial sector ensures that explainability remains the central pillar of Gen AI adoption, thereby preserving stakeholder trust and regulatory confidence as the technology evolves.

