bpi BITS

Financial Services Sector Coordinating Council
for Critical Infrastructure Protection and Homeland Security

Financial Services Sector Coordinating Council (FSSCC)

AI and Explainability in Finance: Explainability Challenges, Practices and Recommendations

January 2026

Table of Contents

Executive Summary

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. At the same time financial institutions are addressing new challenges around 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 such as statistical and machine learning typically provide clear and understandable rationales for their outputs, enabling stakeholder trust, ongoing monitoring and testing to ensure models operate as intended. The emergence of advanced, probabilistic Gen AI algorithms — applied in areas such as fraud detection, cybersecurity, anti-money laundering and customer support — holds significant promise for improving efficiency but also increases the complexity of explainability. Strong governance and risk oversight are essential to maintain stakeholder confidence.

As firms prepare for and implement Gen AI, they are focused on applying foundational principles for decision-making, customer communication and compliance. Maintaining the ability to "explain" decisions outputs requires enhanced approaches to ensure trust and transparency while providing financial services.

To evolve how explainability constructs are applied while using Gen AI, financial institutions are integrating five key disciplines:

- Governance and risk management frameworks that align internal processes and policies to government and industry frameworks like the National Institute of Standards and Technology (NIST) AI Risk Management Framework, Cyber Risk Institute Financial Services AI Risk Management Framework and regulatory guidance and principles
- Data Governance Criteria
- Enhancements to Prompting Guardrails
- Assurance and Testing Methodologies
- Ongoing Risk Monitoring and Outcome Analysis

This paper references various aspects of traditional AI, but focuses more intentionally on Gen AI, underscoring the need for continuing collaboration across the sector, with regulators and third-party providers on how financial institutions can fulfill the core objectives of explainability. The paper includes steps firms should consider to deliver intended and trustworthy outputs, utilize tools effectively and apply guidance to enhance explainable AI and ensure transparency.

As financial institutions adapt and supplement existing governance frameworks, principles and guidance for

deployment of Gen AI, this paper offers a practical resource for a non-technical audiences, business line owners and technology teams to reference as they develop, implement and support AI capabilities.

I. Introduction: The Need for AI Explainability in Finance

Artificial Intelligence, including Gen AI, is transforming finance, creating a need to revisit current methods of explainability to ensure the core tenets of trust, transparency and accountability that exist for traditional AI, are leveraged or updated as the industry unlocks opportunity for innovation, efficiency and improved customer services.

Working with third parties and the regulatory community, the sector aims to ensure AI guidance and compliance requirements remain principles-based and are calibrated to encourage innovation. At the same time, firms must balance current limitations of explainability in Gen AI with ongoing advancements in research around explainability, interpretability, evaluation, etc. This balance is particularly important when considering the broad spectrum of capabilities and use cases, but especially those that help the industry defend itself from cyber attackers, identify fraudulent activity or improve resilience of firms and the sector.

These objectives align with the White House AI Action Plan, which emphasizes advancing U.S. leadership and innovation, balancing these goals with supportive and responsible regulation. The plan calls for research that improves interpretability, control and robustness of models; strengthening critical infrastructure and cybersecurity; promoting resilient, secure-by-design AI systems; and enhancing information sharing and incident response across the public and private sectors.

The goal of this paper is to foster responsible AI by integrating explainability into practice, effectively balancing capabilities and tool deployment, model and non-model performance, products and services with transparency and control. Using outcome-focused oversight that leverages a firm's existing governance and risk management practices, incorporates frameworks such as the Cyber Risk Institute's (CRI) Financial Services AI Risk Management Framework (RMF), the National Institute of Standards and Technology's (NIST) AI Risk Management Framework or principles of model risk management (Federal Reserve Board SR 11-7, OCC, OSFI and the FDIC[1][2][3][4]) where appropriate, along with fair lending, privacy and investor and consumer protection requirements, will best position the sector's use of AI and Gen AI.

As explainability definitions and approaches evolve, implementation is best treated as a dynamic, lifecycle-long objective that is realized through ongoing, risk-based validation, monitoring and outcome analysis. Ongoing dialogue

[1] https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm
[2] https://www.occ.gov/news-issuances/bulletins/2011/bulletin-2011-12.html
[3] https://www.fdic.gov/news/financial-institution-letters/2017/fil17022.html
[4] https://www.osfi-bsif.gc.ca/en/guidance/guidance-library/guideline-e-23-model-risk-management-2027

across the financial sector will help ensure risk-based and technology-neutral frameworks are in place that adapt with advancing innovation.

a. Rise of AI and Generative AI (Gen AI)

Gen AI holds the promise of unlocking significant opportunities, and as financial institutions embrace these technologies, they are also proactively addressing the complexities associated with explainability, transparency, interpretability and trust. By leveraging existing strengths in risk management and governance, institutions are setting a foundation for responsible and transformative Gen AI implementation. Throughout 2024, firms implemented bespoke governance models, data sourcing and control enhancements, and updated oversight practices and testing methodologies for Gen AI. In 2025, use cases entered production and the true benefits of Gen AI started to be measured. As firms develop their use cases, the third-party community is rapidly embedding Gen AI capabilities into their core products including search engines, productivity tools and cybersecurity services. All of these cases are heightening the need for third party transparency and effective oversight policies so firms can adhere to financial regulatory requirements, while improving how the sector uses AI.

As foundation models, large language models, small language models, and firm-sourced data continue to improve in quality and training, Gen AI adoption is expected to accelerate. Over time, Gen AI will enhance efficiency and quality and become integrated into day-to-day operations, such as fraud detection, cybersecurity resilience and broader protections against criminal and adversarial activity across the sector.

b. Trust, Transparency and Accountability

Trust, transparency and accountability, along with effective regulatory supervision, have long been pillars of the financial industry. Through financial institutions' longstanding use of model evaluation and testing, explainability has been foundational for building trust with customers. 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 over time.

The advent of more complex Gen AI algorithms — such as those used in fraud detection, cybersecurity and anti-money laundering — offers immense potential for enhancing efficiency and managing risks. Reaching concrete explainability and full technical measurement, however, remains difficult if not impossible for Gen AI, creating challenges for maintaining compliance and stakeholder trust. Firms apply model-specific explainability approaches and are incorporating alternative and complementary methods (e.g., gradient-based attribution, attention analysis, etc.) to understand AI-generated outputs. This area (noted on p. 22) is rapidly evolving as open-source tools, industry frameworks and academic research continue to emerge.

c. Approaching Explainability Through Existing Risk Practices

To foster improved explainability, financial institutions are integrating five key disciplines: (1) Governance and Risk Management Frameworks, (2) Data Governance, (3) Prompting Guardrails, (4) Assurance and Testing and (5) Ongoing Risk Monitoring and Outcome Analysis.

These risk practices incorporate a series of cross-functional and multi-discipline policies, operating controls and standards of care that firms apply to ensure that Gen AI has tailored and rigorous review. At a high level, this approach includes:

- Data standards, ownership and quality which are addressed further in the "Operationalizing Gen AI" section and is the subject of a more comprehensive FSSCC companion paper on Data Nutrition Labels.[5]
- Tailored validation and documentation for tools, non-models and models, with explainability calibrated to complexity, risk and impact of the implementation.
- Human-in-the-loop reviews for high-stakes or ambiguous decisions (e.g., credit underwriting, fraud alerts).
- Ongoing monitoring for model drift, bias, privacy and anomalous outputs, using outcome analysis to ensure reliability and fact-based metrics are met, with adaptive management to update or retrain as needed.[6] This includes "ongoing monitoring and [a] feedback loop [to] ensure that as use cases, models, data, or regulations change, model and review processes are adjusted and explainability practices are updated and improved."
- Use of prompt engineering, Retrieval Augmented Generation (RAG) and other compensating controls that can set guardrails and help mitigate potential risks of "hallucinations" or inaccuracies in Gen AI outputs and potentially identify malicious activity.

II. Defining Explainability: Concepts and Principles

In financial services, explainability refers to the extent to which one can understand, articulate and justify how an AI or Gen AI model arrives at its decisions or predictions and is a central focus for safe adoption today, and with new and complex use cases. This includes how well one can understand and articulate how and why an AI use case produces a certain output as these results may affect consumers, investors or the operations of financial institutions. Explainability is more than a technical requirement; it is an enabler of trust, accountability and effective decision-making.

Explainability refers not only to understanding the technical mechanisms and data inputs, but also to conveying the outputs and results to a range of stakeholders — including compliance professionals, consumers, regulators and business decision-makers.

The following concepts and principles are utilized to address explainability and are similar to those more broadly included as key principles for explainability in NISTIR 8312.[7]

[5] https://fsscc.org/
[6] https://bpi.com/bpi-comments-on-the-uses-opportunities-and-risks-of-ai-in-financial-services/
[7] NISTIR 8312

Concepts:

- **Explainability:** In financial services, explainability refers to the degree to which the internal processes and decision-making logic of AI can be understood, articulated and communicated.
- **Interpretability:** Includes technical transparency regarding how models interpret data and arrive at specific outcomes, supporting how firms can provide accessible explanations to diverse groups such as regulators, consumers, compliance officers and business leaders.
- **Stakeholder Communication**: Explainability involves tailoring disclosures to meet the informational needs of various stakeholders, ensuring that explanations are appropriately detailed and meaningful for the intended audience.
- **Accountability**: Explainability is crucial for model risk management, regulatory compliance and the ethical deployment of AI systems. It provides a basis for understanding, auditing and reviewing or contesting automated decisions that may have impacts on individuals and financial markets.

Principles:

- **Transparency**: AI decisions and underlying logic should be sufficiently transparent to allow stakeholders to understand the factors influencing individual outcomes and model behavior.
- **Accessibility**: Explanations should use clear language appropriately tailored for technical or non-technical audiences where appropriate, ensuring accessibility for all stakeholder groups.
- **Regulatory Alignment**: Practices around explainability should comply with relevant legal frameworks, including anti-discrimination, consumer protection and data governance regulations in the financial sector.
- **Operationalization:** Explainability should be embedded into the governance structures and risk management practices of financial institutions, allowing for both proactive monitoring and retrospective review of AI-driven decisions.
- **Trust and Fairness**: Providing explainability supports public trust in AI systems, promotes fairness by exposing potential biases and enables stakeholders to challenge or seek redress for adverse outcomes.

Effective explanations should be meaningful, faithful to actual model behavior and candid about limitations — clarifying what can and cannot be explained. Firms' policies and practices are designed to apply these principles, to identify patterns and categorize capabilities and to develop tailored oversight that is "fit for purpose" based on use case, risk assessment and potential impact.

As new services, particularly from third parties, become available, it is most effective to apply these concepts and principles within an agile AI and GenAI lifecycle. This approach is well-suited to the non-deterministic nature of such technologies and aligns with the operational realities of SaaS offerings. The FSSCC work groups see the importance of collaborating with third party providers as this field develops and in defining standards that are

transparent, align with responsible use of AI and comply with consumer and financial industry regulations and guidance.

III. Core Challenges to Explainability: Gen AI — Characteristics of Complexity

a. The Black Box Challenge

Generative AI brings with it inherent complexity, the first of which is the "black box" problem which refers to a system that can be viewed in terms of its inputs and outputs without any knowledge of its internal workings. The inner workings of Gen AI models are opaque, while the contextual and probabilistic output can differ even while answering the same question. This characteristic makes it difficult to interpret or understand Gen AI specifically. These challenges arise due to multiple factors that will be explored in this section.

b. Model Complexity

Gen AI can create original content, such as text, images, video, audio or software codes in response to a user's prompt or request, thus mimicking how humans perceive the world holistically through not only text and voices, but also images and movements. Gen AI leverages natural language processing[8] and deep learning[9] to simulate the complex decision-making power of the human brain.

Gen AI models often involve large quantities of parameters and intricate neural networks to capture intertwined and nonlinear relationships. While there is no consensus on what constitutes "large" in Large Language Models,[10] the LLM model size typically exceeds 6-10 billion parameters. For example, ChatGPT-4 is believed to have hundreds of billions of parameters. The vast number of parameters involved and complex architecture intended to mimic human cognition make it inherently impossible (even for the developer) to understand what factors influence its decision-making processes and contribute to the final outcome. This complexity can also exist within "small" or bespoke language models.

---

[8] NLP definition: What Is NLP (Natural Language Processing)? | IBM
[9] Deep learning definition: https://www.ibm.com/think/topics/deep-learning
[10] See Mitchell, Melanie, MIT Open Encyclopedia of Cognitive Science (2024). "A *large language model* (LLM) is a computational system, typically a deep neural network with a large number of tunable parameters (i.e., weights), that implements a mathematical function called a *language model*. A *language model* (LM), in its most general form, is a probability distribution over possible sequences of words and other elements in a language. ...LMs have been used extensively in many areas of natural language processing, ranging from speech recognition and translation to text generation and chatbots. The neural networks underlying LLMs are trained using broad collections of text typically obtained from websites, digitized books, and other digital resources."

c. Model Non-Determinism

Large language models are the bedrock of Gen AI capabilities. LLMs aim to predict and generate plausible language and other types of content to perform a range of tasks. Autocomplete in word processing is an example of a language model that estimates the *probability* of what comes next in a word or a sentence. Since LLMs learn to predict the next word with probabilities, they are non-deterministic in nature, meaning the same input (prompt) may produce different output.

While the ability of LLMs to generate responses can foster creativity during brainstorming or content creation, it is not the right tool for users who need consistent outputs for decision making, such as instrument pricing. This variability feature also makes traditional validation methods — which depend on predictability and consistency — more challenging to apply.

d. Data Modality

While traditional AI/ML models are trained on structured data that's typically stored in tabular formats such as Excel spreadsheets and relational databases, Gen AI models are trained on a vast quantity of unstructured data inclusive of textual, such as emails, text documents, call transcripts and nontextual, such as images, multimedia files and videos. LLMs are pre-trained on textual datasets and can be considered as an example of unimodal model of one data type. Multi-modal models generally rely on *several* unimodal ones working with multiple data types at the same time to create a multi-faceted description of reality (e.g., a cat that meows), adding to the complexity of Gen AI models.

e. Third Party Risk: Limited Transparency of Vendor Solutions

Many Gen AI solutions leverage pre-trained models developed by third party providers or open-source communities, and as such, access to documentation, training data, developmental details and how the models work is limited. Additionally, for some AI models, enhanced explainability may require revelation of proprietary information, internal details of the system that could be considered a competitive advantage or "intellectual property" of the vendor. In light of existing third-party risk management expectations at financial institutions, the lack of vendor transparency and explainability can add challenges for firms as they evaluate and deploy AI systems.

f. Data Challenges

The accuracy of Gen AI solutions 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 and inaccurate outputs. Data that includes sensitive customer or proprietary information may also introduce privacy risk and compliance issues with data regulations such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA). Ensuring data integrity is thus a critical component of evaluation of conceptual soundness that enhances explainability.

g. Fragmented/Evolving Regulations & Potential Regulatory Burden

A patchwork of current and potential regulation applicable to AI and Gen AI is growing and becoming significantly more difficult for financial institutions, and private industry at large, to navigate. Fragmentation is increasing as states develop their own regulation and financial market guidance alongside insurance regulations, existing or emerging country or regional regulations (e.g., EU AI Act) and new federal legislation being proposed in the U.S. such as the VET Act.[11]

Individually, these can have merit, however, collectively they lead to challenges as financial services firms must adhere to multiple and often inconsistent requirements across geographies and business lines. Attempts in the U.S. to federally preempt, consolidate or rationalize these have not been successful to date, something the White House AI Action Plan prominently notes as an inhibitor to innovation and American advancement. The AI Action Plans call for common regulation at a federal level and sector-based oversight, such as what exists in varying forms across the financial sector, has not taken hold. Legislators have considered over 1,000 AI-related proposals in the first part of 2025; most are at the state vs. federal level. A compilation of potential legislation is included here.[12] [13] [14]

In July of 2025 alone, two proposed measures intended to create common federal standards and reduce the patchwork of state regulations were removed from larger legislative packages. The rapid legislative activity reflects ongoing debate about the role of states, federal oversight, and risk management for emerging AI technologies. It highlights the significant challenges the financial sector and the nation as a whole would experience without consolidated and consistent baseline standards and requirements.

The White House Action Plan prominently points out the need for regulatory harmonization. Without some regulatory consolidation the likelihood exists that inconsistent, incompatible, duplicative and over-burdensome regulation would greatly inhibit the advancement of AI. The potential for overlapping state, federal and market regulations to disrupt advances in AI is prominent for financial institutions, who are also subject to existing financial regulatory requirements and supervision. This imbalance creates a significantly uneven playing field for financial firms, stifling innovation and creating an inability to launch new competitive services as quickly as non-financial firms and fintechs, giving non-financially regulated firms an unfair advantage without the built-in protections consumers and customers rely on. Additionally, the potential for regulation to throttle the use of AI to protect against cyber attackers, fraudsters and scammers could expose the sector to potential operational disruption and expose consumers to unnecessary risk.

[11] https://www.congress.gov/bill/118th-congress/senate-bill/4769#:~:text=S.4769%20%2D%20VET%20Artificial%20Intelligence%20Act
[12] https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states
[13] https://www.brookings.edu/articles/how-different-states-are-approaching-ai/
[14] https://www.rila.org/blog/2025/09/ai-legislation-across-the-states-a-2025-end-of-ses

IV. The Interconnected Roles of Explainability and Interpretability

Explainability and interpretability are often used interchangeably, making the distinction between the two murky while the inter-relationship and co-dependency between these is important to recognize as they serve different yet related purposes. Based on the inter-relationship of these two disciplines, common practice frequently pairs them together. In straightforward terms:

- **Interpretability** refers to the extent to which an expert can understand the internal mechanics of a model. It allows technical experts to predict model performance and inspect internal logic, providing transparency about "how" results are generated, and supporting audit and debugging of risk exposures. Interpretability is a prerequisite that supports explainability.[15]

- **Explainability** translates these mechanics into business-relevant rationales, often referred to as the "why" behind outputs, communicated in a human-understandable way offering non-expert stakeholders' justifications that support trust, accountability, and regulatory scrutiny.[16]

In practice, effective risk management integrates both disciplines, evidencing not just outputs but also the underlying factors that drive model recommendations, thresholds and variances, and the explanations provided to the users and intended audiences. Both disciplines require balancing simplicity and accuracy: interpretable models are more transparent but less powerful, while complex models need additional tools to make outputs explainable without losing capability.

According to IBM,[17] interpretability is about transparency and enables users (often technical experts) to see "how" features are weighted and combined to generate predictions, requiring disclosure of internal operations for trust. In contrast, explainability is about verification — providing justification for the model's outputs, typically after prediction, often in human language for stakeholders or regulators.

Summarized here, and covered later in this paper, tools are utilized at various process points that are a fit for the model architecture and Gen AI use cases. These tools include explainability methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) both of which are well suited for predictive models that are classic machine learning when Gen AI is used for data preprocessing. These tools can document the preprocessing steps and help present black-box models in a human-understandable way.

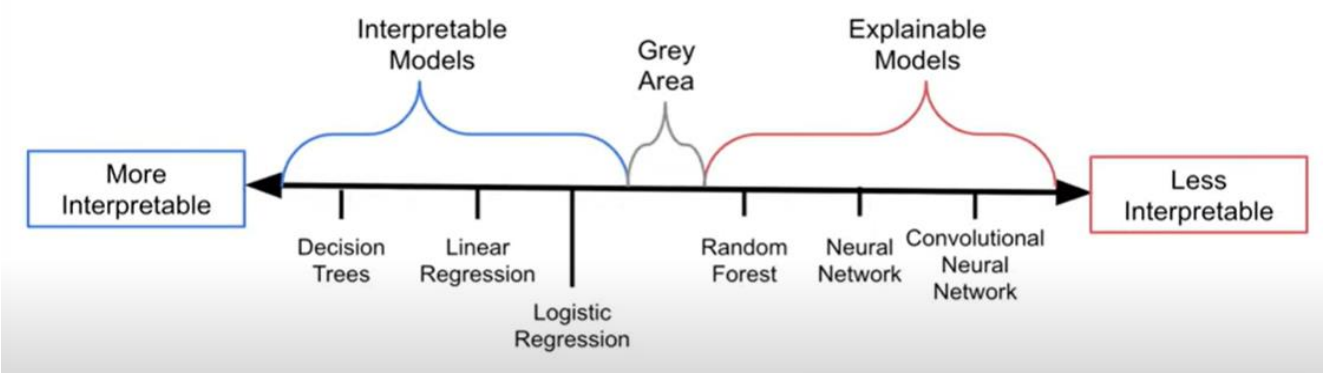
If the predictive model is a transformer or deep generative architecture, neural attribution methods (e.g., Integrated Gradients, DeepSHAP, attention-based methods, etc.) are typically used and modality-specific techniques such as attention rollout for vision transformers and token-level attributions for text, can be a better fit for developing explanations.

---

[15] https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/articles/ai-interpretability-challenge.html
[16] https://sloanreview.mit.edu/article/ai-explainability-how-to-avoid-rubber-stamping-recommendations/
[17] https://www.ibm.com/think/topics/interpretability

A summary chart below illustrates where interpretability and explainability overlap across AI model types.

Figure 1(source unknown)



To try to illustrate the difference in practical use, below are examples of AI — in interpretability and explainability:

- Use Case #1 (Interpretability): In medical diagnostics, clinicians deploy a decision tree model to determine cancer risk. The decision path for each patient is directly visible, with clear splits based on age, biomarkers and test results. With interpretability, researchers and practitioners validate each split and weigh the importance of each feature — ensuring the model aligns with clinical reasoning and catching potential biases or errors before deployment.
- Use Case #2 (Interpretability): In financial services credit scoring models for loan approvals, where transparency is mandated by regulatory requirements, firms may use a decision tree or other interpretable machine learning algorithm to evaluate loan applications. The interpretable model allows compliance officers and auditors to directly observe how features like income, credit history and debt-to-income ratio influence the outcome for each applicant, making it possible to identify and mitigate biases or unfair correlations in decisions.
- Use Case #3 (Explainability, Generative AI): A bank uses a Gen AI model to produce personalized loan approval letters based on client data. The bank must explain to applicants why their loan was denied. Using explainability techniques (e.g., generating natural language explanations and referencing decision thresholds), the bank provides clear, actionable feedback: "Your credit score was 20 points below the approval threshold; assets listed did not meet the minimum requirement." These explanations rely on underlying model interpretability, i.e., the bank's understanding of how inputs relate to outputs, governed internally and communicated externally.

Comparison Chart: Interpretability and Explainability

Following is a representation of interpretability and explainability, comparing at a high level their primary purpose, use and effectiveness for AI and Gen AI.

Aspect	*Interpretability*	*Explainability*
Core Question	"How does the model work internally?"	"Why did the model make this decision?"
Audience	AI experts, developers, engineers	Business users, regulators, general stakeholders
Focus	Transparency of internal logic, structure, parameters	Clarity of decisions, reasons, context
Approach	Inherently interpretable (deterministic) models (e.g., decision trees)	Model-agnostic/post-hoc attribution (e.g., SHAP, LIME)
Usefulness	Debugging, auditing, bias mitigation	User trust, actionable rationale, regulatory compliance
Model Types	Linear regression, logistic regression, simple trees	Deep learning, ensembles, complex/black-box models
Communication Level	Technical, granular	Human-friendly, high-level, logical
Limitation	Often reduces model complexity/accuracy	May oversimplify "how" for the sake of "why"
Dependency	Enables robust explainability	Relies on base model's interpretability

As noted across industries, regulators and financial organizations, balancing risk management, uncertainty and accuracy in Gen AI requires integrating both interpretability and explainability as foundations for trust, compliance, and governance.

As Gen AI models grow more complex and opaque, the challenge lies in ensuring that firms can appropriately justify and govern model outputs — balancing technical insight, business rationale and the variance inherent in generative systems.

V. Capabilities, Use Cases and Context

There are numerous AI use cases deployed and Gen AI capabilities being developed by financial firms (examples as noted in the Appendix) and across a wide spectrum of business lines, service organizations, operating teams and processing centers.

As the use of AI continues to accelerate it amplifies the need to differentiate capabilities, tools, non-models and models into well-defined categories that include a comprehensive understanding of the risk and fit for each use case and processes that calibrate risk and tailor oversight to potential impact. Using pattern recognition disciplines and templates derived from multiple implementations, firms are iteratively streamlining use case review and delivery, while maintaining robust controls and oversight.

Firms have developed cross-functional teams with representatives from legal, privacy, risk management, data and analytics, technology and business and service organizations to assess AI and its use. As firms have evolved their AI programs they are deploying techniques and risk policies to categorize use cases, assess potential benefits and risks (e.g., credit risk, security and privacy risk, etc.), structure decisioning parameters within each, and provide reporting across the enterprise (businesses, operating committees, executive and board committees) to provide transparency and a thorough explanation of the AI being utilized.

As noted above, sample use cases sourced from across the financial industry are included in the Appendix. Each use case presents unique benefits, outputs and risk considerations that fall into four main categories:

- **Operational Efficiency:** Automation and AI-driven insights reduce costs, improve quality and accuracy and free up human resources for higher-impact work. Areas such as routine tasks or software/coding can all be enhanced with AI but also require effective oversight.
- **Risk Management:** Gen AI enhances detection of fraud, cyber threats, and compliance risks, but firms must be cognizant of potential new risks related to explainability, data quality and model drift.
- **Customer Value Proposition and Experience:** Gen AI can enhance customer satisfaction through personalization and service quality and response time improvement, but oversight methods need to be careful to manage privacy, fairness and transparency.
- **Harms:** Firms are using third-party models, generative content, and alternative data sources to improve offerings but must carefully govern these activities to avoid compliance failures or reputational harm.

By tailoring explainability to specific use cases — aligning it with an institution's risk profile, regulatory obligations and business objectives — financial services firms can more readily integrate Gen AI into their operations, providing transparency for different stakeholders — such as internal auditors, customers, business leaders and regulators. This approach can avoid applying overly burdensome requirements for low-risk applications or insufficient rigor for high-risk uses, which could ultimately undermine the effectiveness and trustworthiness of Gen AI implementations.

a. Agentic AI

Financial institutions are evolving from applying AI to discrete well-defined tasks, such as fraud detection or trade execution, toward deploying autonomous agents that automate entire, cross-functional workflows. These agents act as integrated, intelligent systems that can orchestrate complex workflows, analyze unstructured data and support decision-making across front-, middle- and back-office functions. Two current examples are the use of agents for on-boarding new clients and for streamlining payments validation.

In SaaS environments, agents automate functions such as user provisioning, monitoring and workflow orchestration, enabling scalable, adaptive service delivery and continuous integration of new features. Search engines deploy agents to crawl, index, and rank vast amounts of web content, leveraging AI to personalize

results and interpret complex queries. In security models, agents monitor network traffic, detect anomalies, and respond to threats in real time, often coordinating across distributed endpoints to ensure rapid incident response and ongoing risk mitigation.

Across these domains, agents enhance efficiency, adaptability and resilience, but also introduce [...] governance and explainability challenges that require robust oversight and transparent operational controls.[18]

VI. Operationalizing Gen AI — Navigating to Responsible AI Innovation in Finance

To operationalize Gen AI, firms commonly review, modify and augment existing governance and risk practices across every stage of the Gen AI lifecycle, and develop new policies where needed. As noted above, particular implementations differ by firm and use case, therefore, calibrating Gen AI governance and implementation is important as an overly rigid risk management framework could unnecessarily constrain innovation, while insufficient oversight could lead to unintended risks.

A governance model that adapts to technological advancements and business needs will support AI adoption, maintain regulatory confidence and enhance stakeholder trust, enabling U.S. firms to accelerate innovation and drive efficiency. To help understand how to operationalize AI, this paper includes a set of summary approaches and strategies from supporting sources. These include "foundational approaches to ensure explainability in Gen AI implementations prioritizes accuracy and quality throughout the AI lifecycle. This can include validating the technical soundness of model outputs and their business relevance, coupled with the systematic use of explainability metrics such as input traceability, clarity of rationale, and output fidelity. A balanced approach is essential, weighing the trade-offs between predictability and explainability and deploying suitable risk mitigants where necessary to support robust model operations."[19] [20]

a. Foundational Approaches

i. Data Input Governance

A risk-focused approach to data input governance in GenAI emphasizes identifying and mitigating potential vulnerabilities rather than prescribing detailed processes. Financial institutions should prioritize controls that ensure data integrity and reliability — such as mechanisms for validating accuracy and consistency,

[18] https://bpi.com/bpi-comments-on-the-uses-opportunities-and-risks-of-ai-in-financial-services/
[19] https://www.ibm.com/think/topics/explainable-ai
[20] As noted in the Bank for International Settlement Financial Stability Institute Occasional Paper 24, "...given the rapidly changing regulatory landscape and heightened board-level scrutiny, organizations also assess their business risk profiles and compliance obligations in the face of GenAI uncertainties. Effective governance includes maintaining comprehensive support for model decisions, ensuring that explanations provided to regulators, auditors, and clients are clear, transparent, and fit for purpose. This will help build trust and meet the demands of regulatory oversight in critical business areas.
https://www.bis.org/fsi/fsipapers24.pdf

monitoring for drift and maintaining traceability — while recognizing that the primary objective is to reduce risks tied to data provenance, unauthorized use and compliance gaps. Rather than mandating exhaustive step-by-step procedures, risk-based governance encourages ongoing evaluation of exposure to model error, bias propagation and regulatory non-compliance, adapting oversight to evolving threats and business needs. This approach allows firms to align data management practices with their specific risk profiles and regulatory obligations, focusing on outcomes that enhance model trustworthiness and audit readiness.

The Data Nutrition Label (DNL) paper[21] summarized below can be a basis for this effort or highly complementary to existing approaches. The use of automated versioning tools further supports this objective by capturing data changes, making it easier to identify and resolve any potential issues.

ii. Prompting Guardrails with Ongoing Review

One of the most important aspects of Gen AI use is the development of prompting guardrails that are applied consistently, reviewed and updated regularly, and modified to remain consistent with business intent and use case risk profiles. Firms purposefully monitor and guide Gen AI inputs to ensure systems produce reliable, fair and context-appropriate outputs aligned with institutional goals for the given use case. To accomplish this, firms:

- **Design structured prompts**: Develop clear prompt templates that specify acceptable tone, behavior and content boundaries for AI outputs. Templates specify input parameters and contextual guidance to prevent ambiguous or unintended AI behavior.
- **Embed automated filters**: Use technical constraints to detect, flag or block risky, biased or inappropriate results before dissemination. Configure automated detection systems for prohibited topics or harmful content, enabling real-time interventions.
- **Iterate through feedback**: Continuously refine prompt design and instructions based on real-world usage data and user feedback to identify gaps, monitoring outcomes and refine prompt structures to model performance data.
- **Define fine-tuning criteria**: Apply controlled fine-tuning to correct misaligned outputs, update training data and test model revisions against pre-set ethical and performance thresholds before deployment and post-launch. Reviews should include specific performance and fairness benchmarks and user acceptance testing for each model update to confirm compliance with ethical, legal and performance requirements before broad deployment.
- **Integrate stakeholder oversight**: Engage legal, compliance and risk management teams — as well as external experts or vendors — to review prompting standards, assess potential bias and confirm conformity with regulatory and institutional requirements. Firms can also engage third-party

[21] https://fsscc.org/

reviewers or external vendors as needed to independently assess guardrail effectiveness and regulatory alignment.

b. Data Nutrition Labels — A Developing Area

Data Nutrition Labeling, while still a developing practice, is a compelling area that can support the advancement of current data management practices, improve transparency and assist explainability for traditional (deterministic) and Generative (non-deterministic) AI. As Gen AI becomes more integrated into the financial sector, the ability to clearly communicate how these systems produce their results is taking on renewed importance as strong explainability practices help foster trust and support responsible adoption.

Data Nutrition Labels (DNLs) offer a structured way to describe and document the datasets that power AI models. Much like food nutrition labels, they can provide accessible, contextual information — covering data provenance, quality, limitations and potential downsides — that can help teams and stakeholders better understand how underlying data influences AI outcomes intended to enhance clarity and trust. DNLs can complement existing practices and help close gaps between complex data pipelines and the people who need to rely on them by:

- Clearly spotlighting key characteristics, strengths and limitations of datasets, and making intangible data properties explicit.
- Enabling teams to proactively discuss data quality and both positive and negative data attributes.
- Supporting internal governance protocols by documenting key dataset facts.
- Supporting transparency in a way that resonates with both technical and non-technical stakeholders.

DNLs also provide a tangible way to supplement explainability efforts for "black box" Gen AI models, by giving visibility into one of the most influential elements — the data itself.

i. A Practical, Standardized Tool

By offering a structured overview of the dataset "ingredients," DNLs provide a framework to measure how data quality, bias and incompleteness may affect outputs — supporting explainability where full technical deconstruction is not feasible.

DNLs support a consistent, repeatable approach to dataset transparency, and contributing to the confidence among decision-makers, customers and partners; connect technical insights to clear narratives that enhance understanding; strengthen the overall assurance around AI systems without increasing process burden; meet regulatory requirements for data documentation and audit readiness; help communicate clear, explainable narratives to both regulators and the public and enhance responsible use and improve assurance for AI-powered decisions.

ii. How DNL Can Enhance Explainability

- **Improves Dataset Transparency:** DNLs document key data attributes, collection processes, and intended uses, making it easier for stakeholders to understand the data fueling AI models.
- **Identifies Data Limitations and Risks:** Labels highlight dataset gaps, biases, or quality concerns, supporting reviews and risk mitigation efforts.
- **Facilitates Auditability:** Structured data documentation simplifies dataset reviews during regulatory audits, bolstering explainability for compliance teams.
- **Enables Traceability:** DNLs provide enhanced documentation of data sources and their journey through the modeling and development process, supporting increased transparency into how specific inputs contributed to results.
- **Supports Comparability:** Standardized disclosure makes it easier to compare datasets, which can surface divergences, strengths, or exposures across model outputs.
- **Strengthens Accountability:** Explicit documentation of dataset provenance, assumptions, and constraints deepens organizational accountability for model behavior.

In short, DNLs can enhance explainable AI initiatives by adding structured, dataset-level transparency helping to elevate clarity, trust and shared understanding across all audiences.

c. Embed Ongoing Risk Monitoring and Assurance as Core Disciplines

Firms can operationalize explainability through ongoing validation and monitoring "implement[ing] risk-based validation, with the rigor and frequency of validation commensurate with […] complexity and potential impact. Firms conduct periodic assessments to ensure model outcomes remain accurate and fair and perform regular audits for compliance and to demonstrate model accountability to internal and external stakeholders.

In the Model Risk Management (MRM) Guidance and Assurance and Testing sections, firms can describe how ongoing validation (including human-in-the-loop reviews, stress testing, automated monitoring and benchmarking) provides evidence that Gen AI models are functioning as intended and that outputs remain explainable and trustworthy over time.

d. Use Automated and Human Oversight for Model Drift and Anomalies

Ongoing monitoring is necessary to detect when outputs deviate from expected or explainable patterns and involves both automated tools (for real-time anomaly detection, drift monitoring) and human oversight (for interpretability and escalation).

As firms automate, human review may be needed when results are inconsistent or a complex case warrants investigation and remediation of deviations, so that explanations remain accurate and meaningful as models and data evolve.

To apply monitoring consistently, firms implement automated monitoring tools to track shifts in data distributions or model outputs and analyze real-time feedback, user complaints, and operational metrics to spot anomalies or potential biases. They also update or retrain models as needed, ensuring that Gen AI remains accurate, fair and aligned with evolving regulatory expectations.

e. Transparency and Accountability

Institutions should maintain clear records that support traceability and explainability for internal and external stakeholders. Regular monitoring and review of model development and decision processes support compliance, accountability and effective governance without promoting excessive documentation requirements. Oversight should focus on effective governance and outcome monitoring rather than exhaustive documentation, ensuring transparency without imposing unnecessary burden.

f. Tailor Monitoring and Assurance to Context and Risk

An effective method to tailor monitoring and improve explainability is to apply contextual adaptability. Firms tailor explainability practices to intended use, specific business context and risk profile — rather than applying a uniform solution to all AI applications, building a consistent approach that applies rigor and frequency of validation commensurate with the model's complexity, risk and potential impact.

For different audiences this should include context relevant materials and communication tailored and proportionate rather than relying on ever-growing documentation. For example, regulators may require documentation, validation artifacts, and evidence of compliance with laws and guidelines, internal stakeholders will be more interested in technical details for oversight and model management, including data sources, logic, and performance metrics and customers would need to receive clear, jargon-free explanations for decisions affecting them, such as credit denials or flagged transactions, supporting transparency and trust.[22] This tailoring helps calibrate appropriate review of low-risk uses and more intensive oversight for high-impact or high-risk Gen AI applications.

g. Integrate Monitoring and Assurance into Governance

A structured yet flexible governance framework is essential for ensuring accountability, transparency and adaptability in AI implementation. This approach leverages existing risk management practices and integrates ongoing monitoring into Gen AI implementation and management procedures to ensure that variation among

[22] https://home.treasury.gov/system/files/136/Artificial-Intelligence-in-Financial-Services.pdf

Gen AI models meet the objectives of explainability. This integration of ongoing monitoring and assurance is most effective when embedded in enterprise governance structures, ensuring that explainability is tailored to risk and maintained throughout the AI lifecycle and across business lines.

VII. *Frameworks: Standards, Sector AI Risk Management Profiles, Regulatory Guidance*

As there is no single way to measure or define "explainability" what counts as a good explanation can vary by user, use case, risk appetite or regulator. Without common standards, it is hard to compare systems or prove they are trustworthy. We do, however, utilize reference frameworks, which provide a foundation for the use of AI and a basis for explainability. General frameworks such as the NIST Artificial Intelligence Risk Management Framework (AI RMF) and ISO/IEC 4200 can be useful, as is the Cyber Risk Institute's (CRI) development of a Financial Services AI Risk Management Framework, leveraging the NIST AI RMF and other issuances. This effort, similar to CRI's sector specific work on a cybersecurity profile and a cloud profile, will be instrumental in setting financial industry best practices.

a. The NIST Artificial Intelligence Risk Management Framework: Summary

The NIST AI RMF is a voluntary, flexible guide designed to help organizations identify, assess and manage risks associated with AI systems throughout their lifecycle. Developed through extensive consultation, the framework supports trustworthy and responsible AI use by balancing risk mitigation with innovation, promoting ethical practices and technical reliability.

The AI RMF is structured around four primary components — Govern, Map, Measure and Manage. The Govern function establishes oversight, policies and accountability, ensuring effective risk management is embedded into organizational culture. Map focuses on understanding the context and scope of AI systems, recognizing stakeholders and risk sources across the AI lifecycle. Measure involves risk assessment, including technical, ethical, and operational dimensions, using both quantitative and qualitative methods. Manage guides organizations in prioritizing and mitigating risks, including ongoing monitoring and improvement.

These four components form an iterative cycle, enabling organizations to continuously refine AI risk management and align practices with evolving standards and regulatory expectations. The framework's principles — transparency, fairness, accountability and robustness — ensure that AI deployments are not only effective but also aligned with societal and governance values.

b. Cyber Risk Institute (CRI) Financial Services AI Risk Management Framework

The CRI FS AI RMF identifies AI-related risks, aligns them with NIST AI Principles — such as explainability and interpretability — and extends the NIST AI RMF's Functions (govern, map, measure and manage), Categories and Subcategories to include more specific risk control objectives. The FS AI RMF seeks to place AI-based

technologies in the context of existing financial services governance and risk management practices and highlights where these practices may need to be adapted or extended.

The FS AI RMF also introduces scalable implementation guidance based on "AI adoption levels," which consider how deeply AI is embedded within an institution, factoring in technology, business and risk-based use cases.

The NIST AI RMF and, by extension, the FS AI RMF, emphasizes the need for organization-specific and use case-specific definitions for and acceptable levels of the AI Principles of Explainability, Trustworthiness, etc. For example, the NIST AI RMF's "Measure" Function includes a Subcategory (2.9) for the organization to explain, validate, and document AI models and ensure that AI system outputs are interpreted within context to inform responsible use and governance. The CRI FS AI RMF elaborates on this Subcategory objective by providing four more specific control objectives that apply to organizations that are increasing their adoption of AI.

Example practices included in the CRI FS AI RMF that touch on Explainability, and are based on NIST and a number of regulatory and governmental sources, include ensuring that organizations:

- Model Transparency & Validation (MS-2.9.1): Provide clear and structured explanations of AI models, detailing design choices, training data sources, feature importance and decision models. The model validation process should include testing for stability, fairness and performance across diverse conditions, with specific attention to potential biases and limitations. Organizations should employ interpretable models for post hoc explainability techniques where possible and ensure that AI outputs align with intended policies and trustworthy AI principles. As organizations adopt AI, they should consider implementing advanced, automated processes and real-time updates informed by ongoing data analysis.
- Contextualized Output Interpretation (MS-2.9.2): Ensure that AI-generated outputs are interpreted in context, helping users understand how decisions were reached and potential limitations. Decision rationales should be structured to be accessible to all types of teams, users and stakeholders, so that decisions align with the expected use case, audience expertise, and impact considerations. Organizations should integrate human oversight mechanisms where necessary and apply explainability testing methods to validate the clarity and effectiveness of AI-generated explanations. At higher AI adoption levels, these practices should be automated and dynamically integrated into oversight and testing mechanisms.
- Role-Based AI Training (MS-2.9.3): Develop and deliver role-specific training and educational resources to help stakeholders, interpret, validate and responsibly apply AI outputs. Organizations should ensure that training materials incorporate examples of AI decision logic, system constraints, and fallback mechanisms to enable informed decision making. As AI becomes more embedded, organizations should automate the development and delivery of training, incorporating real-time examples of decision logic, constraints, and mechanisms. These processes should be seamlessly integrated into the organization's training and development frameworks.

- Ongoing Governance and Reassessment (MS-2.9.4): Periodically reassess its AI model documentation, validation techniques, and interpretability strategies based on new data, stakeholder feedback, and evolving best practices in responsible AI governance. Updates should be incorporated into governance processes to ensure continued reliability and trustworthiness of AI systems. At more advanced AI adoption stages, reassessment should be automated and embedded into the organization's broader governance and risk management frameworks to ensure transparency, interpretability and accountability.

Organizations should consider these practices as foundational elements of ensuring explainability and interpretability of AI models. The specific methods for how each of these outcome-based control objectives may be achieved are left to the organization and may evolve over time as the business and technology landscape shifts. The FS AI RMF includes other control objectives that also support the principle of explainability but may be secondary to other principes.[23]

c. Technical Standardization Tools to Improve Evaluation, Measurement and Explainability

Evaluation and measurement are receiving significant attention as a core component of the White House AI Action plan, the Office of Science and Technology Policy (OSTP) initiatives, through scholarly research in academic and scientific circles, and in open source and technical development efforts.

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, may be confidential to the provider and may need contractual modifications to be shared, creating barriers to achieving full explainability.

Although at the time of this writing, there is a lack of standardized evaluation measures for Gen AI explainability, there are open-source tools emerging and in use, such as SHAP and LIME, and technology tools that support emerging analytic capabilities and evaluation and measurement frameworks. Gen AI models may also require additional, model-specific explainability approaches, such as attention analysis, gradient-based attribution activation patching and rationale/chain of thought explanations to provide meaningful insights into their outputs. Combining multiple tools to enhance explainability is becoming a common practice.

As Gen AI evolves and expands the capability of these tools (see appendix), the development of new tools will be critical for building trust, supporting regulatory audits and enabling transparent, responsible AI deployment in high-stakes environments.

Example techniques and open-source tools in use today include:

[23] www.cyberriskinstitute.org

- SHAP and LIME are the most widely used for both local and global model explainability supporting a range of model types and use cases. Grad-CAM for example can reveal which regions of an input image or tokens in a text sequence the model prioritized.
- Feature Importance is a technique that determines the contribution of each feature in a dataset to a model's output and answers the question: "what features have the most influence over model's prediction" thus enhancing model explainability.
- Retrieval-Augmented Generation (RAG) retrieves relevant information to generate more accurate and up-to-date responses by combining retrieval-based and generation-based models. RAG is often used to set guardrails and to provide specific data sources or documentation to support its generated response thereby promoting explainability and trust.

Examples of open-source technology tools:

- ELI5 is valued for its simplicity and integration with other tools, making explainability accessible to non-experts.
- InterpretML and AIX360 provide advanced, enterprise-ready solutions with support for multiple explainability techniques, visualization, and regulatory compliance.

d. Regulatory Guidance, Principles and Applicability

MRM principles — model development, model validation, monitoring and outcomes analysis — can be applied to AI, while forms of Gen AI may require alternative approaches. Firms evaluate a model's design, assumptions, alignment with business objectives and logic for their desired use case to find the right fit and use ongoing monitoring and rigorous outcome analysis to enhance reliability. These efforts help firms tailor and calibrate review efforts to high-risk and low-risk use cases (such as simple chatbots or productivity tools), applying more robust oversight to high-risk uses and avoiding unnecessary burdens on low risk uses that don't improve risk identification and control.

Ongoing monitoring: verifies continued performance as business conditions and data evolve. This involves routine checks of model outputs, benchmarking against relevant comparators (including internal and external sources), and adjusting for risks specific to GenAI models. This is especially relevant when using vendor solutions or managing Gen AI-specific risks like output drift and hallucination. Techniques such as drift detection, challenger models, and stability testing help identify emerging weaknesses. Responsive monitoring identifies areas for recalibration as products, data, or external vendor inputs change. Ongoing monitoring helps firms assess "whether a model warrants adjustment, redevelopment, or replacement due to changes in data collection methods, products, environment, shift in user behavior, etc., including changes due to information provided by third-party vendors."[24] The use of benchmarking helps track variances and strengthen confidence in complex or opaque GenAI models, often revealing that greater complexity does not necessarily yield better

[24] SR 11-7 p12

performance and that simpler models may be more effective. Ongoing monitoring can also incorporate new performance metrics tailored to Gen AI's unique risks to ensure sustained reliability and trustworthiness over time.

Outcome analysis: compares a model's outputs to real-world results, using both quantitative and qualitative tools. These tests help surface bias, toxicity, or spurious output — assuring that models perform ethically and as intended, even in complex or emerging use cases. Advanced bias testing frameworks, toxicity classifiers, and synthetic audit data are increasingly used to strengthen this assessment and help ensure outcomes don't inadvertently inherit or amplify inconsistencies or inaccuracies present in the training data, generate fake citations with confidence (e.g., hallucinations) or include toxic content.

Detecting bias, hallucination and toxicity levels and evaluating relevance and accuracy of output is an important part of outcome analysis. Firms can integrate specialized tools — such as bias and toxicity detection APIs — that automatically screen generated content for problematic outputs. For instance, deploying a solution like Perspective API or Google's Vertex AI Model Monitoring allows the firm to detect and flag issues in real time before sharing results with users.

Overall, developing a monitoring and analysis foundation for AI and Gen AI and deploying these effectively not only enhances the accuracy and relevance of Gen AI-generated outputs but also supports stakeholder confidence by ensuring the content meets business objectives, efficiency goals, ethical use and regulatory standards.

VIII. Conclusion: Recommendations to Accelerate Innovation, Manage Risk and Foster Gen AI Use in the Financial Industry

Explainability is an important obligation for AI and Gen AI use in the financial sector — a critical component for building trust and transparency for consumers and business leaders, and necessary to satisfy regulatory requirements both within the financial services sector and outside.

Firms are adapting and evolving governance processes and "treating explainability as a *dynamic, lifecycle-long objective* that is [...] realized through ongoing, risk-based validation, monitoring, and transparent documentation." [25] The implementation of these approaches is becoming a best practice as AI and Gen AI specifically, migrates to being a "general use" tool for business.

As firms categorize use cases and stakeholder needs, tailor monitoring and assurance to risk, and deploy practices that incorporate regular validation, drift detection, human-in-the-loop engagement and proportionate

[25] https://bpi.com/wp-content/uploads/2024/04/Navigating-Artificial-Intelligence-in-Banking.pdf

documentation, the sector can take full advantage of AI and Gen AI. This method ensures that Gen AI systems in finance remain transparent, accountable and trustworthy — even as models, data, and risks evolve.

On a go forward basis it's important that the sector work across government agencies and with regulators, third parties and research organizations to meet the dual mandate of driving efficiency and innovation while balancing risk. Areas of focus include:

- Deploying an effective lifecycle that includes well defined use cases, data input standards such as those identified in the Data Nutrition Labels paper, testing and assurance processes and human-in-the-loop decisioning.
- Collaborating across the firm to ensure functional roles (e.g., business owners, data and privacy, governance and risk management, technology, etc.) have input into AI use cases and continue to modify, develop and adhere to governance and risk policies.
- Using standards-based frameworks such as the CRI FS AI Risk Management Framework (informed by the NIST AI Risk Management Framework) to organize how best to implement AI responsibly.
- Maintaining a focus on transparency and accountability by applying policies across classes of use cases, conducting periodic outcome reviews and tailoring communications to explain how a model arrived at its decisions on a stakeholder-by-stakeholder basis.
- Minimizing the "black box" nature of GenAI by applying model-specific methods (gradient-based attribution, activation patching, and rationale/chain-of-thought), iteratively fine-tuning models and services and using evaluation models, prompting guardrails, benchmarks, and measurement frameworks to assess outputs.
- Improving observability and monitoring with data attribution, human-in-the-loop review of workflows, ongoing monitoring and drift detection, and retrieval/knowledge transparency (provenance, citations, auditability).
- Establishing forums for collaborative dialog between industry, government and regulators (e.g., FSSCC, trade association and information sharing organizations) so that risk management guidance is informed by industry input and emerging technological trends and ensuring that supervisory practices are consistent and evolve alongside innovation.
- Fostering regulatory harmonization and a consistent national approach via agencies such as the White House Office of Science and Technology Policy to develop practical and consistent ground rules that level the playing field across industries, promoting fairness and competition.
- Working with service providers and third-party technology firms to deploy evaluation and explainability tools that help satisfy transparency and regulatory obligations for the safe deployment of AI and Gen AI products.

The FSSCC working group would like to thank the many participants who helped develop this paper. Your input from across the sector and from third party service providers was instrumental in advancing this work as the use of AI and Gen AI expanded in the financial industry. The substantial contributions from firms, associations, standards bodies and regulators, and your willingness to re-factor this paper based on what was learned from initial implementations have been invaluable.

As the FSSCC continues its work on AI and Gen AI to help shape the use of technology, promote effective regulation and take advantage of AI innovation, we look to develop responsible, trustworthy and transparent adoption of AI that improves services and products while advancing cybersecurity capabilities, reducing fraud and fostering a more resilient sector.

APPENDIX AND SUPPLEMENTAL INFORMATION

Artificial Intelligence and Generative AI Use Cases

Below is a compilation of example use cases from multiple sources. As firms continue defining new use cases and opportunities to deploy AI and Gen AI we anticipate this list will expand significantly:

Traditional AI Use Cases

- Fraud Detection: Machine learning models are used for credit/debit cards, digital payments, and account opening fraud detection. For example, graph databases map customer-merchant interactions and payment flows, reducing fraud rates, losses, and improving customer experience.
- Fraud Detection & AML: AI models for fraud detection and anti-money laundering are governed by rigorous validation, human-in-the-loop review, and ongoing monitoring. Institutions document model logic, data sources, and performance, ensuring that flagged transactions can be explained to auditors, regulators, and customers.
- AML/Transaction Monitoring: AI identifies transaction anomalies and assigns risk scores for review. ML is used to link customer records for more efficient KYC checks.
- Credit Decisioning: ML models (e.g., gradient-boosted decision trees) support credit line increases/decreases, underwriting, and collections, using both traditional and alternative data. These models are validated for fairness and explainability, especially for adverse action notices under ECOA and Regulation B.
- Credit Decisioning: For AI-driven credit underwriting, governance frameworks require clear documentation and the ability to provide adverse action notices, as mandated by fair lending laws. This ensures that decisions are explainable to consumers and regulators, and that models can be audited for fairness and bias.
- Customer Servicing: AI is used for "next best action" recommendations, personalized offers, and natural language processing to resolve customer inquiries.
- Cybersecurity: AI-driven systems monitor network traffic, detect anomalies, and automate responses to spam/phishing, improving threat response and reducing analyst workload[17].
- Analytics: AI platforms process news and internal metrics for risk management and business decision support, with human verification before action.
- Back Office Functions: Includes financial reporting, knowledge management, and employee productivity enhancements (e.g., chatbots for policy questions).

Generative AI Use Cases

- Software Development: Gen AI assists coding, code review, and documentation, with human oversight.
- Risk Management: Gen AI reviews third-party cybersecurity responses, flags gaps, and drafts vendor communications, reducing review time and focusing human attention on high-risk areas.
- Cybersecurity: LLMs with chatbots support "white hat" hacking and vulnerability identification, referencing large knowledge bases.
- Adverse Media Screening: Gen AI and NLP distribute news to AML analysts, flagging risk signals and event context.
- Product Development: Gen AI creates thematic indices by scanning news for keywords, supporting new product offerings.
- Marketing and Personalization: Gen AI generates marketing content and personalizes offers, with human review for compliance and brand consistency.
- Customer Service Enhancement: Gen AI summarizes support conversations, identifies common issues, and improves chatbot responses using RAG and purpose-built LLMs.[4]
- Customer Service: Gen AI chatbots and virtual assistants are subject to prompt engineering, output monitoring, and escalation procedures. Governance includes regular review of model outputs for accuracy and appropriateness, with logs and documentation supporting explainability for both internal review and customer inquiries.
- Internal Knowledge Management: Gen AI is used for content extraction, summarization, search, and audio transcription/translation[5]

Emerging and Exploratory Use Cases

- Regulatory Compliance and Suptech: AI tools support regulators in market surveillance, risk identification, text analysis, and automating supervisory processes.
- Financial Inclusion: AI-driven credit scoring models using alternative data expand access to credit for underserved populations.[6]
- Research and Advocacy: AI analyzes consumer complaints, studies financial abuse risks, and raises awareness of scams (e.g., robocall analysis).

Impactful Example Use Cases from the CB Insights report - August 2025

- AI-Powered Virtual Assistants for Customer Support: Institutions like Truist and Wells Fargo have deployed generative AI assistants to manage millions of client interactions, automating routine responses and reducing the need for human intervention. These platforms significantly boost efficiency, cut response times, and improve scalability—but present risks related to accuracy, customer experience, and data privacy.

- Contract Intelligence and Loan Documentation Automation: Some major banks have implemented AI systems to accelerate loan processing and contract management. These solutions help streamline compliance and reduce transaction cycle times but demand robust governance for auditability and regulatory assurance.
- Trade Surveillance and Anomaly Detection: AI is now used in trade monitoring to automate detection of suspicious activities or market manipulation. This yields quicker risk identification but requires explainability and transparency, especially for regulatory reporting.
- Claims Management and Underwriting in Insurance: Generative AI chatbots assist with claims documentation, triage, and personalized policy recommendations, improving customer service and lowering costs. Risks include model hallucinations and the potential for incorrect claims evaluations.
- Wealth Management Advisor Copilots and Investment Analytics: Asset managers deploy generative AI to synthesize research, generate personalized portfolio summaries, and enhance client decision-making. While this drives hyper-personalization, it amplifies the challenge of maintaining consistency, transparency, and robust compliance practices.
- Agentic Commerce and Automated Payments: Mastercard, Visa, and PayPal are piloting agentic AI that enables autonomous execution of transactions and payments. These solutions promise convenience but carry significant regulatory, security, and explainability considerations as AI agents automate financial decision-making.

These examples illustrate both the breadth of Gen AI applications and the contextual importance of tailored explainability and risk management within financial services. Expanding on these areas will highlight real-world impact, stakeholder benefit, and governance challenges central to your narrative

Use Case Examples from Major SaaS providers

- Customer Relationship Management (CRM): Salesforce delivers a cloud-based CRM platform that enables organizations to manage sales pipelines, automate marketing campaigns, and provide customer support from any device. AI-powered features include predictive sales analytics, automated lead scoring, and personalized customer engagement.
- Enterprise Resource Planning (ERP): SAP S/4HANA Cloud and Oracle NetSuite offer comprehensive ERP solutions as SaaS, integrating finance, procurement, supply chain, and HR processes. These platforms support real-time analytics, regulatory compliance, and global business management, reducing IT overhead and accelerating digital transformation.
- Collaboration and Productivity Suites: Microsoft 365 and Google Workspace provide cloud-based productivity tools, including email, document editing, video conferencing, and shared storage. These platforms leverage AI for features like smart scheduling, automated meeting transcriptions, and real-time language translation.

- Cybersecurity and Identity Management: Okta and Microsoft Entra ID (formerly Azure AD) deliver identity and access management as SaaS, enabling secure single sign-on, multi-factor authentication, and automated user provisioning across cloud and on-premises applications. These services use AI agents to detect suspicious activity and automate threat responses.

- Contact Center and Customer Support: Genesys Cloud CX and Zendesk provide SaaS-based contact center solutions, using AI-powered virtual agents to route inquiries, automate responses, and analyze customer sentiment across channels. These platforms help enterprises scale support, improve customer satisfaction, and gain actionable insights from interactions.

- Data Analytics and Business Intelligence: Snowflake and Tableau Cloud offer SaaS platforms for scalable data warehousing, analytics, and visualization. Organizations use these tools to integrate disparate data sources, run advanced queries, and share interactive dashboards globally, with AI-driven insights and anomaly detection.

These use cases demonstrate how leading SaaS providers are leveraging cloud delivery models, embedded AI agents, and robust security controls to drive efficiency, scalability, and innovation for enterprise customers.

Explainability Tools

As of the date of this paper, some of the most promising explainability tools in use today—widely adopted across industries and especially valued in regulated sectors like finance include the following:

Tool Name	Key Features & Strengths	Best For
SHAP	Uses Shapley values from game theory to attribute predictions to input features. Provides both local and global explanations, works across model types (linear, tree-based, neural networks), and offers intuitive visualizations.	Detailed feature importance, model-agnostic interpretability, regulatory reporting.
LIME	Generates local, model-agnostic explanations by approximating the model around a prediction with an interpretable surrogate. Useful for explaining individual predictions and debugging black-box models.	Local explanations for individual cases, debugging, compliance documentation.
ELI5	Python library that visualizes and explains model predictions, showing feature weights and importances for a variety of model types. Integrates with LIME and permutation importance and is user-friendly for beginners.	Simple, human-readable explanations, model debugging, education.

InterpretML	**Microsoft's** toolkit supporting both inherently interpretable (glass-box) models and black-box explainers (like SHAP and LIME). Offers global and local explanations, what-if analysis, and visualizations.	Multiple interpretability techniques, comparison of methods, enterprise use.
AI Explainability 360 (AIX360)	**IBM's comprehensive open-source** toolkit with a wide range of algorithms (including SHAP, LIME, Anchors), visualization tools, and bias detection. Designed for enterprise and regulatory environments.	Comprehensive, enterprise-grade explainability, fairness analysis, regulated industries.

Additional Sources and References:

"Navigating Artificial Intelligence in Banking", BPI/BITS Staff, April 2024 https://bpi.com/wp-content/uploads/2024/04/Navigating-Artificial-Intelligence-in-Banking.pdf

"Continuous Risk Management Strategies for AI Advancements", Megha Thakkar, April 17, 2025 https://www.scrut.io/post/continuous-risk-management-for-ai-advancements

"Top 10 AI Risk Governance Tools for Regulating Generative AI in Enterprises" April 3, 2025, updated October 16, 2025 https://www.cloudnuro.ai/blog/top-10-ai-risk-governance-tools-for-regulating-generative-ai-in-enterprises-2025-guide

"BPI Comments on the Uses, Opportunities and Risks of AI in Financial Services, BPI Staff, August 12, 2024 https://bpi.com/bpi-comments-on-the-uses-opportunities-and-risks-of-ai-in-financial-services

"U.S. Department of the Treasury Request for Information on Uses, Opportunities, and Risks of Artificial Intelligence in the Financial Services Sector, TREAS-DO-2024-0011 Ben Tecmire, Dr. Rachel Schutt, August 12, 2024- https://www.blackrock.com/corporate/literature/whitepaper/blackrock-response-to-treasury-ai-rfi.pdf

"Artificial Intelligence in Financial Services – Report on the Uses, Opportunities and Risk of Artificial Intelligence in the Financial Services Sector", Staff Paper, December 2024 https://home.treasury.gov/system/files/136/Artificial-Intelligence-in-Financial-Services.pdf

"Treasury's Post 2024 RFI Report on AI in Financial Services – Uses, Opportunities, and Risks" Charu A. Chandrasekhar, Avi Gesser, Erez Liebermann, Gregory J. Lyons, January 24th, 2025 https://www.debevoise.com/insights/publications/2025/01/treasurys-post-2024-rfi-report-on-ai-in-financial

"Explainable AI (XAI) in Cybersecurity: Understanding AI-based Security Decisions" Staff Blog, July 8, 2024 https://akitra.com/explainable-ai-xai-in-cybersecurity/

"What Is Explainability? Palo Alto Networks, Sourced October 2025 https://www.paloaltonetworks.com/cyberpedia/ai-explainability

"Explainable AI (XAI) in Security Applications" Rishika Patel, September 24, 2024. https://aithority.com/machine-learning/explainable-ai-xai-in-security-applications/

" The 5 Best AI (XAI) Explainable Tools in 2025" Data World (from Service Now), Sourced September, 2025 https://data.world/resources/compare/explainable-ai-tools/

" Explainable AI: 5 Open-Source Tools You Should Know" Gilad David Maayan, February 21, 2024 https://tdan.com/explainable-ai-5-open-source-tools-you-should-know/31589

"Top 10 Tools for Achieving AI Transparency and Explainability in Enterprise Settings" Staff Webpage, June 20, 2025 https://superagi.com/top-10-tools-for-achieving-ai-transparency-and-explainability-in-enterprise-settings/