REPORT

TECHNOLOGY

CFA Institute Research & Policy Center

Explainable AI in Finance: Addressing the Needs of Diverse Stakeholders

Cheryll-Ann Wilson, PhD, CFA

Executive Summary

No, no! The adventures first. . . . [E]xplanations take such a dreadful time.

—The Gryphon in Lewis Carroll's *Alice's Adventures in Wonderland*

Decision-making systems orchestrate our world, powered by machine learning (ML) systems based on artificial intelligence (AI). These AI-based systems help underwriters and credit analysts to assess risk, portfolio managers to optimize security allocation, and individuals to select investment and insurance products. As the digital economy grows, so does the need for immense computing power. This power comes at a cost, however: Systems based on deep learning algorithms in particular can become so complex that even their developers cannot fully explain how these systems generate decisions.[1] This, in essence, is the "black-box problem," which makes it difficult to trust an AI system's decisions, assess model fairness, and meet regulatory demands. Consequences include actual or perceived discrimination against protected consumer groups and violation of fair lending rules.

This problem has led to the consideration of various proposed solutions—the most well known being explainable AI (XAI) technologies—to create a cognitive bridge between human and machine. XAI refers to AI and ML techniques, or capabilities, that seek to provide human-understandable justifications for the AI-generated output. Implicit in explainable AI is the question "explainable to whom?" In fact, defining "whom" (or the user group) is essential to determining *how* the data are collected, *what* data can be collected, and the most effective way of describing the reason behind an action. This report focuses on the human behind human–machine collaboration. The objective is to generate

[1]Deep learning algorithms are described in greater detail in Wilson (2025, p. 6).

CONTENTS

discussion on the best way to support the needs of diverse groups of AI users. As such, this report explores the role of XAI in modern finance, highlighting its applications, benefits, and challenges, with insights from recent studies and industry practices. It presents a detailed analysis of the explainability needs of six stakeholder groups, the majority of which are nontechnical users. The analysis includes matching their needs with their job responsibilities and assessing the most relevant XAI methods. Finally, the report reviews some alternative approaches to XAI—evaluative AI and neurosymbolic AI.

With its focus on AI explainability, this study represents a deeper analysis of transparency and explainability issues raised in earlier CFA Institute works. These publications include "Ethics and Artificial Intelligence in Investment Management" (Preece 2022) and "Creating Value from Big Data in the Investment Management Process" (Wilson 2025).

Key Takeaways

1. **The Need for AI Explainability in Finance**
   - *Credit scoring and lending:* Deep learning models can provide more detailed assessments by using alternative data (e.g., credit card transactions, social media), but they require explainability to ensure fairness, transparency, and regulatory compliance.
   - *Investment and portfolio management:* AI can enhance financial analysis, asset allocation, and risk management by detecting patterns in large datasets to improve modeling and decision making, but lack of explainability and model "hallucinations" can lead to misinformed decisions and financial losses.

- *Insurance:* AI can speed up underwriting, boost fraud detection, and enhance customer service, but its use raises concerns about unintended bias and discrimination created through correlations with sensitive personal attributes. Examples of nonpersonal characteristics that may indirectly correlate with protected attributes include zip codes as proxies for socioeconomic status or ethnicity, as well as purchasing history for gender or ethnicity.
- *Regulatory challenges:* AI-driven systems present oversight difficulties caused by limited transparency in data sources and decision-making logic.

2. **Explainability Techniques**

This report categorizes XAI methods into two main types:

- *Ante-hoc (built-in explainability) models:*
  - Designed to be inherently interpretable (e.g., decision trees, linear regression, rule-based systems)
  - Provide *global explainability*, offering transparency in how a model works overall
  - Useful for regulatory and risk management applications where interpretability is prioritized over predictive accuracy
- *Post-hoc (after-the-fact explainability) models:*
  - Applied to black-box models (e.g., deep learning, ensemble methods) to generate explanations *after* predictions are made
  - Examples:
    - Feature attribution methods (SHAP, LIME): Determine which input factors influenced an AI decision
    - Visual explanations: Heatmaps, partial dependence plots, and attention maps to illustrate AI reasoning
    - Counterfactual explanations: Explain how a decision could have changed under different circumstances (e.g., "If income were $5,000 higher, the loan would be approved")
    - Rule-based and simplification approaches: Approximate black-box models with more interpretable versions

3. **XAI Applications**

This report addresses the following key examples. This list should not be construed as exhaustive, however.

- *Credit scoring and lending:* XAI methods, such as SHAP and LIME, can help financial institutions justify loan approvals or denials.
- *Algorithmic trading and investment strategies:* Visual techniques, such as heatmaps, can help traders understand how models generate buy/sell signals.

- *Fraud detection and anti–money laundering (AML):* Feature attribution techniques are used to improve the interpretability of fraud detection models.
- *Regulatory compliance and risk management:* Regulators require clear explanations for AI-driven financial decisions, ensuring accountability and fairness.

4. **Key Challenges in Implementing XAI**
   - *Technical challenges:*
     - Lack of standardized evaluation metrics: No universal benchmarks exist to assess the quality of AI explanations, leading to inconsistent evaluations.
     - Real-time decision-making constraints: Delivering instant, understandable explanations during fast-paced transactions remains difficult.
   - *Regulatory challenges:*
     - Privacy risks: Detailed explanations can unintentionally reveal sensitive personal or financial data.
     - Absence of universal explainability standards: Differing regional regulations (e.g., EU versus US regulations) create compliance challenges for firms that operate internationally.
   - *User experience challenges:*
     - Overreliance on AI explanations (algorithmic appreciation): Users often trust AI outputs without critical evaluation, leading to confirmation bias.
     - Limited user-friendly tools: Most XAI tools are built for technical users, with a lack of accessible interfaces for business users, regulators, and customers.
5. **Alternative Approaches to XAI**

   Beyond standard XAI frameworks, the report explores the following:

   - *Evaluative AI:* Focuses on hypothesis-driven decision making rather than direct AI recommendations, promoting human engagement
   - *Neurosymbolic AI:* Integrates rule-based reasoning with deep learning to improve interpretability while retaining predictive power

   XAI presents a transformative opportunity for financial institutions to enhance transparency, regulatory compliance, and trust in AI-driven decision making. Although challenges such as overreliance on explanations, privacy risks, and model complexity persist, strategic adoption of XAI can help financial firms navigate these obstacles effectively. By developing standardized frameworks, tailoring explanations to stakeholders, balancing interpretability with performance, and ensuring privacy protection, financial institutions can use XAI to the extent of its potential while maintaining ethical and responsible AI practices. Future research should focus on developing hybrid models that balance accuracy with interpretability, creating standardized benchmarks for evaluating XAI methods, and improving computational efficiency in real-time financial applications.

Introduction

The integration of AI into the field of finance has transformed decision-making processes across various domains, including risk management, credit assessment, algorithmic trading, and fraud detection. The opacity of some AI models, however, presents challenges for regulatory compliance and supervision, interpretability, and stakeholder trust. Explainable AI (XAI) techniques have emerged to address these challenges by providing transparency and interpretability in financial decision making. This report reviews the role of XAI to support the explanation needs of diverse stakeholders in the finance sector, many of whom are nontechnical users of these AI models. The review incorporates different financial applications, comparing the effectiveness of various explainability methods, including feature attribution techniques such as SHAP and LIME, rule-based models, counterfactual explanations, and visual explanation techniques. The analysis highlights the potential trade-offs between accuracy and interpretability, the regulatory implications of using XAI, and future research directions for advancing explainability in financial AI models.

In 2024, CFA Institute conducted a multimethodological study of AI and big data use in investment professionals' workflows (Wilson 2025). Investment professionals across a broad swath of regions and occupational categories cited the complexity and opacity of AI models (the "black box" or explainability issue) as the second-biggest impediment to greater AI adoption in organizations. The black-box issue was also a key point of concern among the C-suite executives, practitioners, and regulators who participated in roundtable sessions. The issue of model opacity also was raised in an earlier CFA Institute study on ethics and AI (Preece 2022).

Numerous researchers have acknowledged the need for human-understandable AI systems that customize explanations based on the specific user's needs, knowledge, and goals (Brennen 2020; Gerlings, Shollo, and Constantiou 2021; Ribera and Lapedriza 2019; Tomsett, Braines, Harborne, Preece, and Chakraborty 2018). Notwithstanding, the focus of many studies in the past 50 years has been explainability for model developers (Miller, Howe, and Sonenberg 2017).

Given the ubiquitous concern and ever-increasing complexity of AI systems, a deep dive into the explainability needs of diverse stakeholders and purported solutions offers useful insights. This report seeks to frame the relevant issues around the tools and approaches created to deal with AI model opacity from the perspective of the human at the center of the operations. This study is designed to help finance and investment professionals to navigate real-world ethical challenges arising from the use of AI and big data technologies. In addition, it aims to stimulate discussion among core stakeholders—including practitioners, C-suite executives, policymakers, and regulators—around effective ways to augment the collaboration between humans and machines and embed the human-in-the-loop principle into AI system design.

This report extends the existing body of CFA Institute work on AI, big data, and machine learning.[2]

The Need for AI Explainability in Finance

In the highly regulated financial sector, actions based on decision-making AI systems can have significant implications for consumers, businesses, and the economy. AI in finance encompasses a wide range of applications, including investment research and analysis, portfolio management, trading, risk management, lending, and customer service—all of which require a high level of trust and transparency (Cao 2023). The following examples from the banking, investment, insurance, and regulation sectors illustrate the real-world consequences of AI inscrutability, also known as the black-box problem.

Credit Scoring and Lending

Traditional credit scoring models often rely on a limited set of variables, which may fail to capture the full financial picture of an individual. Deep learning AI models can incorporate alternative data sources, such as transaction history and social media activity, to provide a more comprehensive assessment. The complexity of these models, however, requires explainability to ensure fair and unbiased lending decisions. Kuiper, van den Berg, van der Burgt, and Leijnen (2022) highlighted the importance of XAI in consumer credit and mortgage lending, where transparency is crucial for both regulatory compliance and consumer trust.

Investment and Portfolio Management

Financial professionals are increasingly engaging AI to support investment decision making, using sophisticated algorithms to analyze vast datasets, identify patterns, and enhance asset allocation strategies. Although these tools have the potential to improve efficiency, reduce human bias, and optimize returns, significant risks remain with respect to their use. Many AI models function as opaque systems, making it difficult to interpret or validate their recommendations. This lack of transparency can lead to misguided investment decisions, particularly when users place unwarranted trust in the technology. Notably, in 2019, Bloomberg reported a case in which an entrepreneur based in Hong Kong SAR lost $20 million after relying on advice from an AI-powered investment platform (Fearn 2024). Risks are further amplified by the potential for "hallucinations," in which AI systems, especially those based on large language models, generate plausible but factually incorrect outputs caused by limitations in their training data. Such errors could mislead investors and result in substantial financial losses.

A similar concern has arisen in the private credit sector. Private credit has comparable characteristics to a bridge loan: Financing is typically short term (i.e., between nine months and five years) with double-digit variable interest

[2]CFA Institute publications related to technology, big data, and AI are available on the Research and Policy Center website at https://rpc.cfainstitute.org/themes/technology.

rates (Olson 2025). Nonbank financial institutions typically provide private credit to middle-market firms (i.e., firms with annual revenues between $10 million and $1 billion; see Cai and Haque 2024). The largest investors in private credit funds tend to be pension funds, insurance companies, family offices, sovereign wealth funds, and high-net-worth individuals (Cai and Haque 2024). To keep abreast of heightened competition, private credit firms have begun to use generative AI and machine learning technologies to help dealmakers vet potential investments and develop sophisticated investment strategies (Taylor 2024). Investment professionals are increasingly concerned, however, about potential biases in AI datasets that could adversely affect decision making and potentially lead to unreasonable outcomes.

Insurance

AI offers many insurance-centric applications, notably in underwriting, fraud detection, and customer service (Wilson Drakes 2021). Until fairly recently, insurers depended on expert judgments and simple rule-based heuristics to make critical predictions. With the aid of deep learning applications, underwriters can combine past experience with data from digital maps and high-resolution satellite and drone imagery to quickly assess risks in property and casualty insurance (Karapiperis 2019).

In fraud detection, AI-based systems are superior to conventional statistical predictive models because they can quickly scan enormous amounts of data in different formats, such as claims adjusters' handwritten notes, repair estimates, and claimants' social media accounts (Karapiperis 2019). ML algorithms can be trained to discover new (and new variations of) fraud patterns by inspecting data anomalies that may go undetected by human investigators. Potential benefits to insurance firms are financial (reduced losses from fraud), operational (more targeted deployment of investigative resources), and reputational (avoiding adversarial customer interactions by not challenging legitimate claims [Karapiperis 2019]).

Although analyzing data at a very granular level may help an insurance company to more efficiently align its pricing with its risk assessment, this practice could lead also to discriminatory outcomes. Insurers are not allowed to consider data on sensitive characteristics such as ethnicity, religion, or gender, but machine learning algorithms may use geographical data or other individual attributes (Wilson Drakes 2021). An AI system may thus generate outcomes that implicitly correlate with those sensitive characteristics (Karapiperis 2019), in the process creating a "structural elaboration"—that is, a situation in which the actual result works against the ideal (Joseph, Ocasio, and McDonnell 2014).

Regulation

AI-based ML systems raise several regulatory issues. Regulators may struggle to inspect and supervise firms if data types, sources, or decision-making processes are unclear, making it hard to assess financial risks accurately. Companies that

use these complex models without fully understanding them can face unintended consequences. Furthermore, proprietary concerns may prevent firms from offering full system explainability to regulators.

The following section explores the regulatory landscape as it relates to AI explainability.

The Regulatory Landscape

Emerging regulations are shaping AI explainability requirements, particularly in such high-risk domains as finance, health care, and legal decision making. In the European Union, for example, the Artificial Intelligence Act (EU AI Act) mandates transparency and human oversight for high-risk AI systems, including detailed documentation on training data and evaluation methods. As noted by Liesenfeld and Dingemanse (2024), however, exemptions for open-source models in the EU AI Act could weaken transparency standards, allowing providers to evade the requirements for detailed disclosure of training data and fine-tuning methods. This regulatory gap poses a challenge to the accountability of generative AI systems and their explainability.

Similarly, the US "Blueprint for an AI Bill of Rights" (White House OSTP 2022) emphasizes the right to an explanation for AI-driven decisions, particularly in sensitive applications such as credit scoring and hiring. This emphasis aligns with the Equal Credit Opportunity Act, which requires that AI-driven credit decisions provide "specific and accurate reasons" for adverse outcomes. As Moreno (2024) highlights, however, the effectiveness of explainability tools such as explainable AI is contested because of epistemic risks and inconsistencies in model interpretation. Epistemic risk is defined as "the risk of being wrong" (Biddle 2016) when selecting hypotheses, methodologies, assumptions, datasets, or policies.

XAI models face a unique epistemic risk because they are "models of models": They attempt to explain the workings of opaque AI systems, but their own methodological choices shape what aspects of the black-box AI they reveal (Moreno 2024). As a result, different XAI approaches can yield divergent explanations for the same decision, creating uncertainty about the reliability of post-hoc AI explanations. Moreover, studies have shown that AI providers may strategically select the most convenient XAI framework to obscure biases or ensure regulatory approval, raising concerns about fairness and accountability (Krishna, Han, Gu, Wu, Jabhari, and Lakkaraju 2022).

Other emerging regulations, such as Canada's Directive on Automated Decision-Making and OECD guidelines, reinforce the need for explainability but without clearly defining enforceable standards. OECD (2023) warns that a lack of explainability in generative AI models constitutes a major risk, particularly in finance, where opaque AI models complicate trading and investment strategies and credit risk assessments. Regulators are increasingly advocating

for preemptive fairness assessments and mandatory transparency reports to ensure AI systems remain interpretable.

Although explainability requirements are central to AI regulations, their implementation remains complex. The divergence in XAI methods, legal loopholes for open-source AI, and strategic opacity by providers challenge regulatory efforts. Future policies should seek to close existing loopholes in this arena, as well as establish standardized explainability metrics and robust certification frameworks to ensure that AI decisions are interpretable, accountable, and aligned with ethical guidelines.

Key Concepts Related to Explainability

The Cambridge Dictionary's definition of the term "explanation" aligns with the human-centered focus of this report: "the details or reasons that someone gives to make something clear or easy to understand." This definition, used in this report, underscores two essential attributes of an explainable system: clarity and understanding. It also implicitly places the onus on the explainer—in this case, the AI model—to provide clarity in a manner that is understandable (Ribeiro, Singh, and Guestrin 2016; Gilpin, Bau, Yuan, Bajwa, Specter, and Kagal 2018; Rudin 2019) to the human requiring the explanation (i.e., the explainee).

The concept of explanation serves as the baseline for the other essential concepts behind explainable AI, such as transparency, fairness, and accountability. The following are other important features of explanations:

- *Explanations are dynamic*: They are a process that may take several interactions between explainer and explainee in order to produce a satisfactory outcome (Mueller, Hoffman, Clancey, Emrey, and Klein 2019).
- *Explanations are contextual*: Not everything needs to be explained. The need for an explanation is a function of the specific user's information needs relative to the event (Mueller et al. 2019; Miller 2019).
- *Explanations are social in nature*: Knowledge is transferred as part of an interaction between the explainer and explainee (Miller 2019). The power of this attribute lies also in its generality; this interactive quality can be used to explain both human and technical actions.
- *Explanations are not a "one size fits all" phenomenon*: The diversity of backgrounds and explanation needs requires flexibility on the part of the explainer to meet the context.
- *Explanations are often triggered by violations of expectation* (Hoffman, Klein, and Mueller 2018): For instance, a customer may seek a reason for being denied a loan, given that he or she has been granted similar facilities in the past.

Exhibit 1 depicts the two-dimensional nature of explainability. The following discussion elaborates on the key features of and differences between ante-hoc, or "built-in," explainability and post-hoc, or "after-the-fact," explainability.

Exhibit 1. Two Sides of Explainability

Category	Ante-Hoc Models	Post-Hoc Models
Coverage	Global explainability	Local explainability
Predictability	Potentially limited predictability	Potentially limited interpretability
Integration	Internal: part of the model structure	External: add-on feature
Focus	Wide focus: captures inner workings of the model as a whole	Narrow focus: able to explain specific decisions/predictions
Model dependency	Model specific	Model agnostic
Perspective	Technical perspective	Social perspective
Explanation type	Process-based explanations	Outcome-based explanations
Target audience (broadly)	Model developers	End users

Ante-Hoc Models

Ante-hoc explainable models are designed to be interpretable from the outset, meaning they are inherently transparent and understandable by humans. Ante-hoc models often use simpler, more interpretable algorithms (vis-à-vis post-hoc models), such as decision trees, linear regression models, or rule-based systems. The goal is to make the model's decision-making process transparent and easy to follow, which can be particularly important in fields such as finance where trust and accountability are crucial. Note that when used in the context of explainable systems, transparency almost invariably refers to AI model transparency. Rudin (2019) advocates for inherently interpretable models instead of trying to explain opaque AI models.

Global explainability refers to how the decision-making process is made transparent (Guidotti, Monreale, Ruggieri, Turini, Giannotti, and Pedreschi 2019; van den Berg and Kuiper 2020; Wanner, Herm, and Janiesch 2020). Global explainability provides transparency at the level of (1) the entire model (i.e., simulatability), (2) individual components of the model (i.e., decomposability), and (3) the training algorithm (i.e., algorithmic transparency; see Lipton 2018; Wanner et al. 2020). To achieve algorithmic transparency, the model must demonstrate an ability to produce predictable outcomes, even when used on new datasets (Lipton 2018). Global explainability is normally used to explain or interpret simpler models.

In contrast, local or outcome-based explanations of AI systems (Information Commissioner's Office and The Alan Turing Institute 2020) facilitate an *instance-based* view of interpretability only: in other words, the reason(s) for a *specific* prediction, decision, or outcome (Guidotti et al. 2019; Yeo, Van Der Heever, Mao, Cambria, Satapathy, and Mengaldo 2025). Therefore, local explainability focuses on clarifying the reasoning behind a particular algorithmically generated

outcome in plain, easily understandable, everyday language (Information Commissioner's Office and The Alan Turing Institute 2020). The subsequent subsection, "Post-Hoc Explainability," discusses local explainability in more detail.

The trade-off for increased transparency in AI models often comes at the cost of reduced predictability. The primary goal of transparency is to make the internal structure and functioning of the model more understandable, including the datasets used during training (Tomsett et al. 2018; Lipton 2018; Preece 2018). This focus on revealing the underlying *processes* of ML models is particularly relevant to system developers and other technically proficient stakeholders (Lipton 2018; Wanner et al. 2020). Such process-based explanations aim to demonstrate that robust governance procedures and best practices are being followed throughout the design and deployment of AI systems (Information Commissioner's Office and The Alan Turing Institute 2020).

This transparency/predictability trade-off is closely connected with the concept of the bias–variance trade-off. In this context, *bias* refers to the assumptions or constraints built into a model—such as those required by linear regression, which assumes normality, homoskedasticity, and linearity. Models with higher bias, such as linear models, are generally more transparent and often exhibit lower variance in their predictions when applied to new data. Their predictive performance may be limited by the structural assumptions imposed on them, however. Conversely, models with low or no bias—such as nonlinear models and deep learning architectures—tend to offer greater predictive power but are more sensitive to fluctuations in the data, leading to higher variance in their outputs.

Although the bias–variance trade-off is traditionally used to understand such issues as overfitting and underfitting, it is also relevant here as a lens through which to understand the relationship between model transparency and predictive performance.

Post-Hoc Explainability

Although model transparency may be geared toward system engineers, post-hoc explanations focus on end users who might not be system professionals (Lipton 2018; Preece, Harborne, Braines, Tomsett, and Chakraborty 2018). In this case, the XAI is said to be extrinsically explainable (Yeo et al. 2025) because it operates more like an "add-on" or external feature to the AI model (Preece et al. 2018). Unlike ante-hoc models, which are built with interpretability as a core feature, post-hoc explanations attempt to explain the decisions of already-trained models (Gunning and Aha 2019).

Achieving global explainability in practice is difficult with such complex machine learning systems as the multilayered deep neural networks. Such systems are inherently hard to interpret and prone to failure if asked to extrapolate (Páez 2019; Elton 2020). As a result, local explanations tend to be preferred for such systems (Montavon, Samek, and Müller 2018; van den Berg and Kuiper 2020; Wanner et al. 2020).

Local explainability may enable post-hoc explanations of the AI model's results through textual explanations (in natural language), visualizations, or classifications based on similar examples (Lipton 2018). These techniques are usually viewed as XAI from a social perspective, aimed at strengthening trust and use of an AI model (Miller 2019) through interactions between different stakeholders at various stages in model development (Hong, Hullman, and Bertini 2020).

Irrespective of which fork in the explainability road is selected—ante-hoc or post-hoc—consensus exists among researchers (as well as among regulators) on the necessity to explain AI-based outcomes in contextually appropriate and human-understandable terms.

Explainability Methods in Finance

Various XAI techniques have been developed to enhance interpretability in financial AI applications. As discussed in the preceding section, these methods can be broadly categorized into post-hoc explainability methods and inherently interpretable models (Weber, Carl, and Hinz 2024). Some methods used to improve the explainability of deep learning models in the context of finance include visual explanation, feature attribution, feature relevance, explanation by simplification, and explanation by example (Yeo et al. 2025; Černevičienė and Kabašinskas 2024). Each method provides a different way to enhance transparency in AI-driven financial decision making.

Exhibits 2 through 7 illustrate how some of the XAI techniques discussed in the following subsections can be applied to machine learning models commonly used in fundamental factor investing. For this purpose, James Tait, Affiliate Researcher at CFA Institute Research and Policy Center, selected XGBoost because it is one of the most popular and powerful gradient-boosting algorithms. He trained an XGBRegressor model, a gradient-boosting model from the XGBoost library. The model emulates a factor-investing strategy, a common use case in portfolio construction and performance attribution. Tait used a dataset retrieved from Bloomberg LP that contains normalized exposures to six Barra style factors across 218 stocks over a 100-month period starting in February 2008.

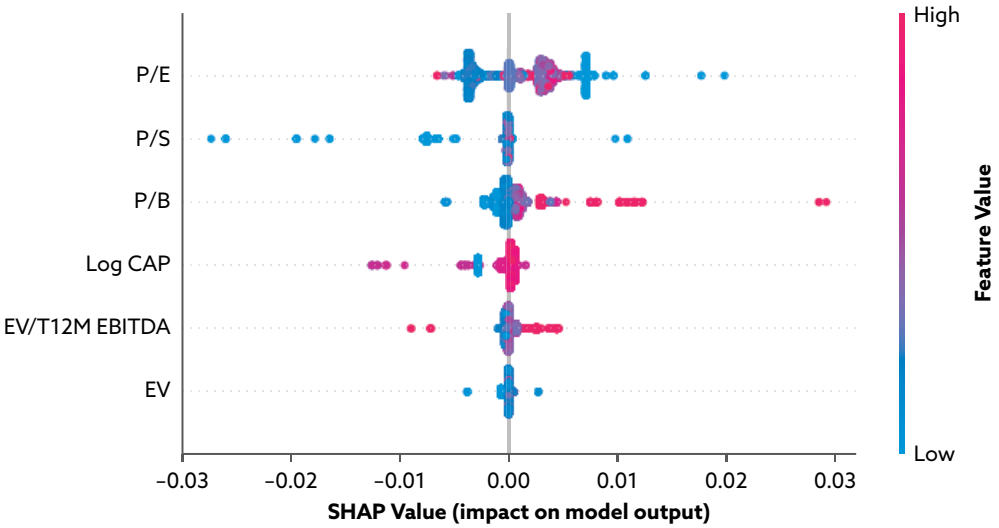
For a detailed demonstration (including code) of how explainability tools can be used in financial modeling, visit the CFA Institute Research and Policy Center GitHub repository.[3]

Feature Attribution

Feature attribution methods are primarily used to determine how much each input feature contributes to a specific AI model prediction. These methods are particularly useful for local interpretability, meaning they explain individual predictions rather than general trends across a dataset.

[3]https://github.com/CFA-Institute-RPC/.

Exhibit 2. SHAP Plot



Notes: P/S is price-to-share ratio; P/B is price-to-book value ratio; Log CAP is the logarithm of market capitalization; EV/T12M EBITDA is enterprise value to trailing 12-month earnings before interest, taxes, depreciation, and amortization; and EV is enterprise value. This chart can be found at the CFA Institute Research and Policy Center "Explainable-AI-in-Finance" GitHub repository: https://github.com/CFA-Institute-RPC/Explainable-AI-In-Finance.

Source: Bloomberg LP.

One of the most widely used feature attribution techniques is SHAP (SHapley Additive exPlanations), which is based on cooperative game theory and assigns importance scores to each input feature. In finance, SHAP is used extensively in credit risk assessment, fraud detection, and economic forecasting (Yeo et al. 2025). For example, when assessing a loan application, SHAP can reveal whether an applicant's credit score, debt-to-income ratio, or recent delinquencies had the most impact on the model's decision. Similarly, in fraud detection, SHAP helps identify which transaction characteristics, such as unusual spending patterns or geolocation mismatches, contributed to a flagged transaction (Le, Nauta, Nguyen, Pathak, Schlötterer, and Seifert 2023).

In high-frequency trading, decisions are made in milliseconds. XAI models such as SHAP are increasingly used to explain trade executions, ensuring alignment with investment strategies and regulatory requirements (Sudjianto and Zhang 2021). **Exhibit 2** illustrates a SHAP plot derived from the use case outlined in the preceding section. Each point on the SHAP summary plot represents a single prediction for a single data point, showing how much a specific feature contributed to that prediction—both in magnitude and direction. For example, high values in the price-to-book ratio (P/B, associated with growth stocks) typically led to increased monthly return predictions over the period from 2008 to 2016.

Feature Relevance

Feature relevance (or importance) methods, in contrast, assess the global importance of features across an entire model rather than explaining individual predictions. These methods are typically used to determine which features have the most influence on a financial model's overall decision-making process.

Linear regression—arguably the simplest and most interpretable machine learning method—is often overlooked in discussions of explainable AI. In linear regression, the explanatory power of each input feature is made transparent through its associated coefficient, which directly indicates the proportion of variance in the target variable accounted for by that feature. Although frequently considered a basic statistical tool, this clarity of interpretation makes linear regression a textbook case of model transparency.

Beyond linear models, more complex techniques are used for feature relevance in nonlinear and ensemble models. One widely used approach is permutation feature importance (PFI), in which the values of a particular feature are randomly shuffled and the resulting drop in model performance is measured. A larger decrease in accuracy indicates greater importance of a feature. Another method is Gini importance, often applied in tree-based models, which ranks features according to how much they contribute to reducing impurity or variance at decision nodes throughout the model.

Additionally, models such as XGBoost offer built-in feature importance metrics that fall into this same category of feature relevance techniques. The more a feature is used to make key decisions with decision trees, the higher its relative importance. Metrics include Gain (the average improvement in the model's loss function—such as log-loss—from splits using a feature); Cover (the average number of samples a feature split affects, reflecting how broadly a feature influences the dataset); and Weight (how often a feature is used for splitting). Although Gain is most closely aligned with a feature's predictive contribution, all three measures provide useful perspectives for ranking and interpreting feature relevance within gradient-boosted decision trees (Chen and Guestrin 2016).

Exhibit 3 presents a comparative summary of the PFI, Gini, and XGBoost feature importance methods. **Exhibit 4** illustrates the use of the Cover, Gain, and Weight metrics using the XGBoost feature importance tool. The price-to-earnings ratio (P/E) has the highest importance score in the example for each of the three metrics.

In finance, feature relevance techniques are crucial for portfolio management, risk assessment, and fraud prevention:

- *Portfolio optimization:* Feature relevance analysis can determine which macroeconomic indicators (e.g., interest rates, GDP growth, inflation) have the greatest impact on different asset classes, sectors, and securities.
- *Risk management:* Feature relevance techniques, such as PFI and Gini importance, help identify key financial ratios that influence bankruptcy prediction models, allowing banks to adjust lending policies accordingly.
- *Fraud prevention:* By ranking the most influential features in fraud detection models, financial institutions can refine their monitoring systems to focus on the most predictive signals, such as transaction velocity, account age, and behavioral deviations.

Exhibit 3. Comparison of Feature Importance Methods

	Permutation Feature Importance	Gini Importance	XGBoost Feature Importance
Type	Model-agnostic, post-hoc	Model-specific (tree-based)	Model-specific
Computation	Measures performance drop when a feature's values are shuffled	Calculates total reduction in Gini impurity from splits using the feature	Evaluates features based on Gain, Cover, and Weight metrics
Interpretability	High; directly reflects impact on model performance	Moderate; based on impurity reduction	Varies; Gain is most interpretable
Pros	Applicable to any model; Reflects true feature impact	Efficient computation; Provides insight into feature usage	Offers multiple perspectives; Integrated into XGBoost
Cons	Computationally intensive; Affected by feature correlations	Biased towards features with more levels or continuous variables; Not applicable to non-tree models	Specific to XGBoost; May require careful interpretation
Best Use Case	When model-agnostic interpretability is needed	For quick insights in tree-based models	For detailed analysis within XGBoost models

Exhibit 4. Feature Importance: Cover, Gain, and Weight

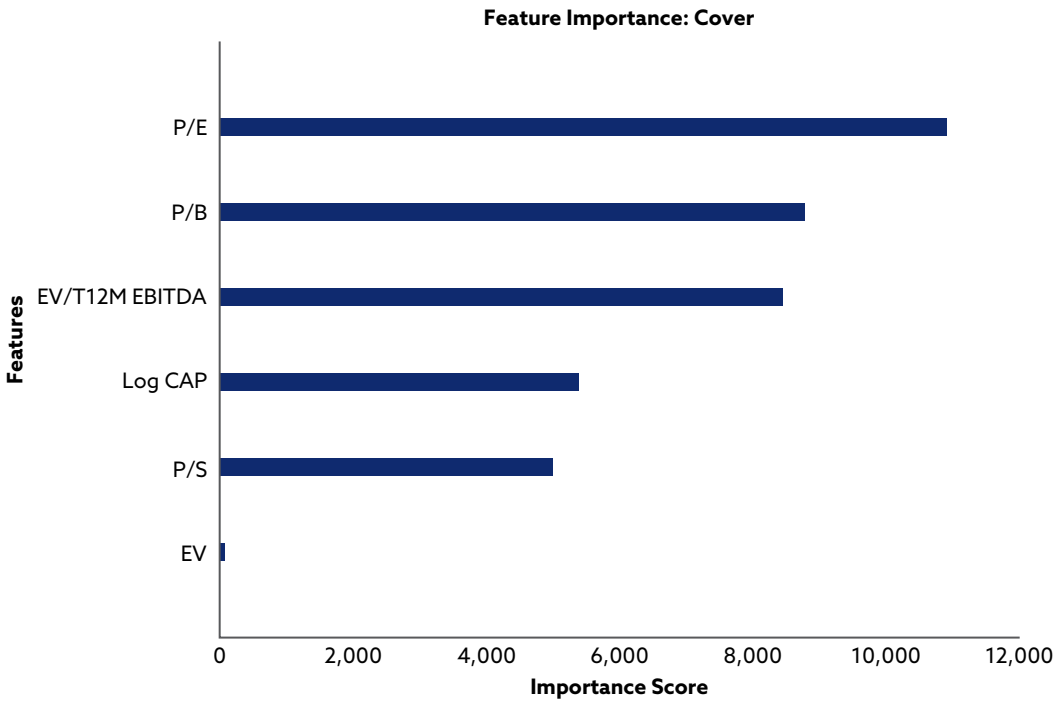
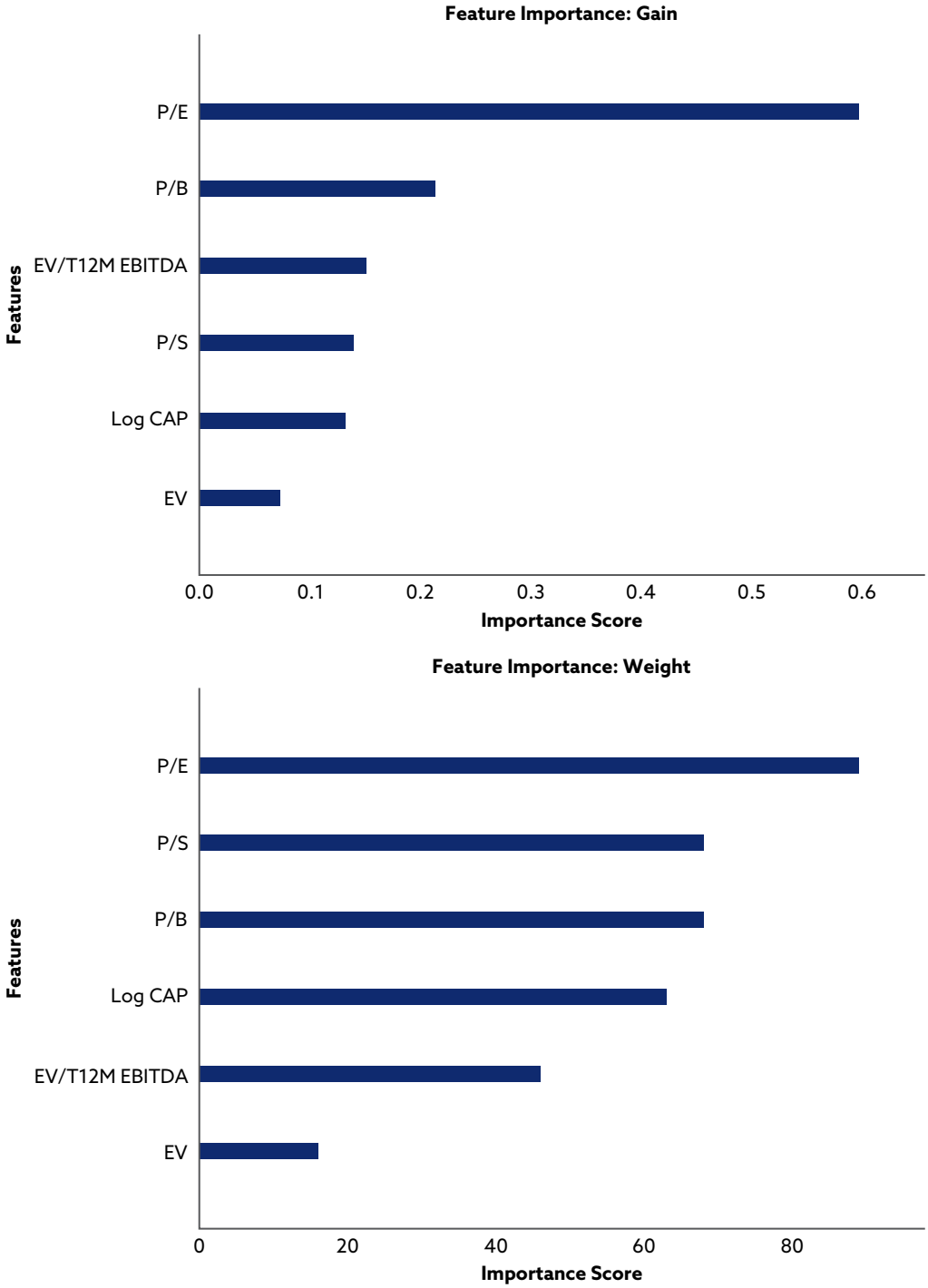


Exhibit 4. Feature Importance: Cover, Gain, and Weight *(continued)*



Note: This chart can be found at the CFA Institute Research and Policy Center "Explainable-AI-in-Finance" GitHub repository: https://github.com/CFA-Institute-RPC/Explainable-AI-In-Finance.

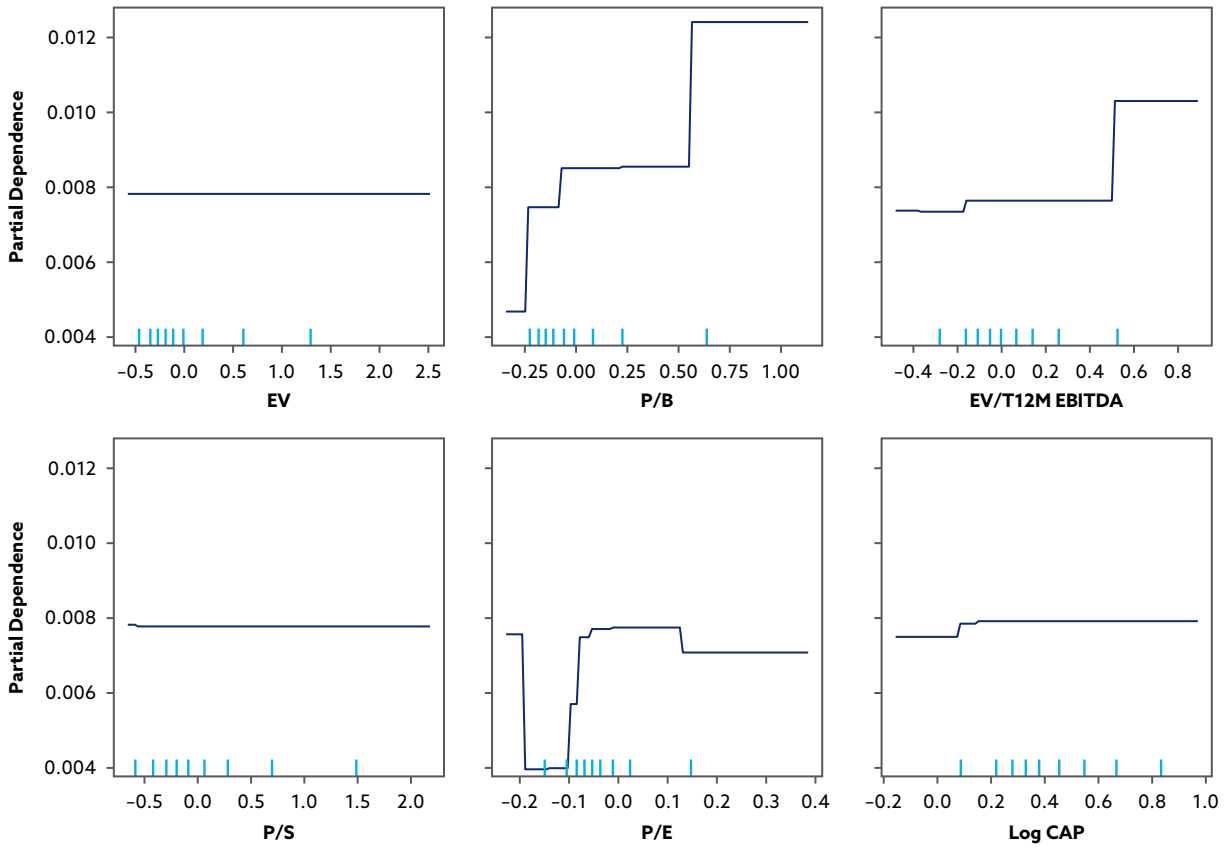
Source: Bloomberg LP.

Visual Explanation

Visual explanation techniques generate graphical representations of AI model decision making. For instance, individual conditional expectation (ICE) plots and partial dependence plots (PDPs) provide visualizations of the effects of feature variations on model predictions (Weber, Carl, and Hinz 2024). These methods are used in finance to audit credit risk models, predict stock movements, and enhance portfolio management strategies by highlighting model behavior and key feature interactions.

In **Exhibit 5**, the PDPs illustrate how changing the value of one feature affects the predictions of the model, while holding all other features constant. For example, the horizontal lines in the price-to-share ratio (P/S) and enterprise value (EV) plots indicate that changing the values of these features has no impact on the model's predictions. In contrast, the step changes in the

Exhibit 5. Partial Dependence Plots



Note: This chart can be found at the CFA Institute Research and Policy Center "Explainable-AI-in-Finance" GitHub repository: https://github.com/CFA-Institute-RPC/Explainable-AI-In-Finance.

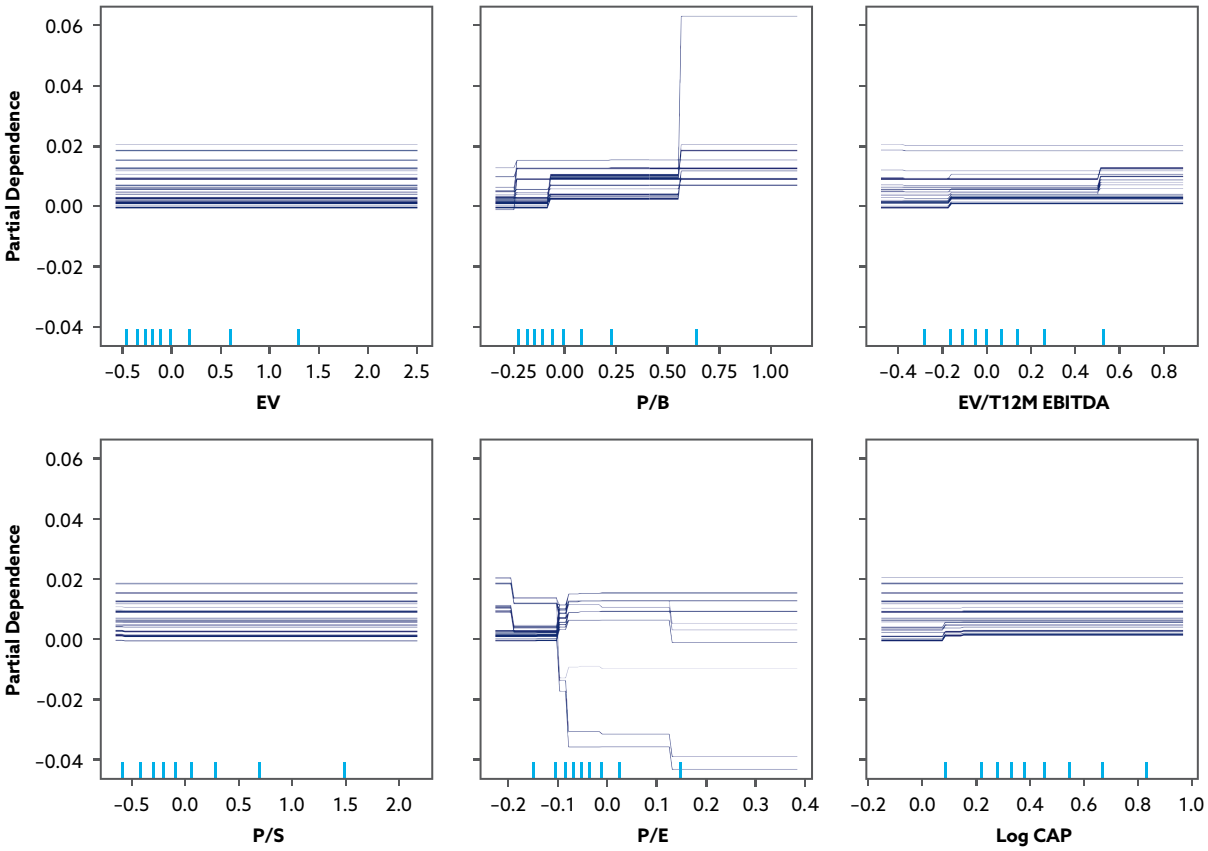
Source: Bloomberg LP.

P/B feature indicate that increasing P/B increases the monthly returns predicted by the model. These observations are similar to those in the SHAP summary; the difference is that each PDP shows the global impact of each feature on predictions, while SHAP shows the effect of each feature on individual predictions.

Additionally, rule-based models and visualization techniques such as ICE plots offer transparency into how AI identifies regulatory breaches (van der Waa, Nieuwburg, Cremers, and Neerincx 2021). The ICE plots in **Exhibit 6** show how modifying the value of one feature affects the predicted monthly return for each individual data point. This scenario also contrasts with PDPs' global perspective.

Counterfactual explanations and visual techniques such as heatmaps assist algorithmic traders in understanding how models generate buy/sell signals (Černevičienė and Kabašinskas 2024). Counterfactual explanations describe the minimal changes needed in an input to obtain a different outcome.

Exhibit 6. Individual Conditional Expectation Plots



Note: This chart can be found at the CFA Institute Research and Policy Center "Explainable-AI-in-Finance" GitHub repository: https://github.com/CFA-Institute-RPC/Explainable-AI-In-Finance.

Source: Bloomberg LP.

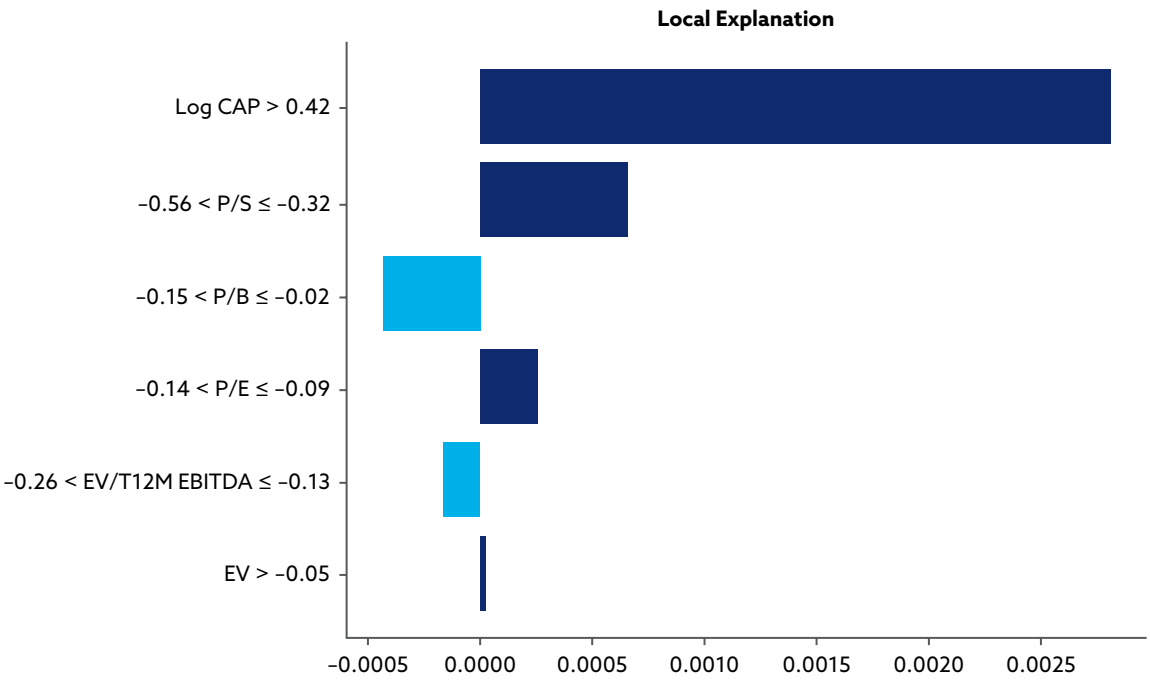
Explanation by Simplification

Explanation-by-simplification techniques approximate complex AI models with simpler, interpretable ones. LIME (local interpretable model-agnostic explanations) is a widely used method that builds locally linear models to approximate predictions from complex models (Yeo et al. 2025). This method is frequently applied in credit scoring, lending decisions, and financial risk modeling to ensure that model predictions align with regulatory and ethical standards (Černevičienė and Kabašinskas 2024). **Exhibit 7** depicts how LIME is used to explain the output of the XGBoost model. In the LIME plot, the four dark blue bars represent the features that positively contribute to the predicted monthly return. For example, the feature Log CAP > 0.42 contributed ≈+0.0030 to the predicted return.

Explanation by Example

Explanation-by-example methods provide interpretability by comparing cases with similar instances. These techniques include case-based reasoning and prototype selection, which are useful in anomaly detection, credit evaluation, and financial fraud detection. In finance, such methods help regulators and auditors understand why a particular decision was made by comparing it with historically similar cases (Yeo et al. 2025).

Exhibit 7. LIME Plot



Note: This chart can be found at the CFA Institute Research and Policy Center "Explainable-AI-in-Finance" GitHub repository: https://github.com/CFA-Institute-RPC/Explainable-AI-In-Finance.

Source: Bloomberg LP.

Together, these XAI methods form a toolkit for interpreting such models in financial contexts, enabling practitioners to validate model behavior, ensure regulatory compliance, and build trust with stakeholders. Although no single method offers a complete explanation, their combination helps bridge the gap between predictive performance and interpretability. **Exhibit 8** provides a comprehensive summary of these methods and their applications in the finance sector.

Explainable to Whom?

As AI becomes a cornerstone of financial decision making, ensuring transparency and trust in AI-driven systems is no longer optional; it is a necessity. Different financial stakeholders, from regulators to financial professionals and customers, require explainability tailored to their unique needs. For example, regulators demand compliance, auditability, and fairness, ensuring that AI models meet legal and ethical standards. Risk managers prioritize model reliability, robustness to stress testing, and risk transparency to mitigate uncertainties in lending and investment strategies. Meanwhile, data scientists and AI developers need scalable and efficient XAI models that balance interpretability with performance, ensuring that complex algorithms remain both explainable and effective.

For traders, investment analysts, and portfolio managers, real-time decision making and market trend interpretation are critical, requiring visual and feature-attribution-based explanations to support AI-driven investment strategies. On the customer side, loan applicants need clear decisions and explanations to understand how financial decisions affect their creditworthiness, and investors need to understand how an AI model is used to support the stated investment objectives of the investment product or mandate to ensure suitability. Internal auditors and compliance teams further ensure that AI models align with organizational governance and risk management frameworks. By implementing such XAI techniques as SHAP, LIME, counterfactuals, and rule-based models, financial firms seek to enhance transparency, build trust, and ensure that AI-driven decisions are not only powerful but also accountable, fair, and understandable to all stakeholders.

Exhibit 9 outlines the requisite needs for AI explainability among various stakeholders, as well as how these needs relate to their responsibilities.

Challenges of Implementing XAI in Finance

Despite its many benefits, implementing XAI in the finance sector presents several challenges. One key issue is the overreliance on AI-generated explanations. Many stakeholders, including nontechnical users, tend to trust AI explanations without critically evaluating their validity. This phenomenon, known as algorithmic appreciation, was discussed in two CFA Institute reports: "Pensions in the Age of Artificial Intelligence" (Hayman 2024) and

Exhibit 8. XAI Methods Used in Finance

Aspect	Feature Attribution	Feature Relevance	Visual Explanation	Explanation by Simplification	Explanation by Example
Definition	Assigns importance scores to individual features for a specific prediction	Measures the overall importance of a feature across the entire model	Uses visual representations to highlight which parts of the input are most influential	Simplifies a complex model by approximating it with a more interpretable model	Explains predictions by providing similar examples from a dataset
Scope	Local (instance specific)	Global (feature importance across the model)	Local (for an instance) or global (for an entire dataset)	Global (simplification applies to the entire model)	Local (case-based explanations)
Techniques used	Uses Shapley values (SHAP), integrated gradients, or perturbation methods to determine feature importance per prediction	Assesses model sensitivity to feature changes through PFI, feature ablation, or Gini importance	Uses heatmaps, ICE plots, or PDPs	Fits a surrogate model (e.g., LIME, decision trees, linear regression) to mimic the original black-box model	Uses *k*-nearest neighbors, prototypes, counterfactuals, or case-based reasoning to retrieve similar cases from a dataset to justify a decision
Model dependency	Often model agnostic but can be model specific (e.g., gradient-based methods)	Can be model specific (e.g., decision tree feature importance) or model agnostic (e.g., permutation importance)	Model agnostic or model specific (e.g., convolutional neural network attention maps)	Model agnostic (fits to any black-box model)	Model agnostic (relies on dataset similarity)
Output	A score per feature for each instance	A ranking or average of scores across the dataset	A visual heatmap or highlighted regions of the input	A simplified model that approximates the black-box model	A set of similar examples supporting a decision
Use case	Explaining why a specific prediction was made (e.g., why a loan was denied)	Understanding overall feature importance for model transparency	Explaining image classification, natural language processing (NLP) models, or black-box predictions using visual cues	Making AI decisions understandable by approximating them with an interpretable model	Justifying decisions based on historical cases (e.g., legal decisions, medical diagnosis)

Exhibit 9. Stakeholder Needs from XAI Models in Finance

Stakeholder Group	Key Responsibilities	XAI Needs	Description
Regulators and compliance officers	Ensure **compliance** with financial regulations (e.g., General Data Protection Regulation, Basel III, EU AI Act)	Regulatory justification	Financial firms must provide justification for loan denials, credit scoring, and AML decisions
	Monitor AI-driven financial systems for **fairness and transparency**	Transparency and interpretability	AI decisions must be explainable and traceable
	Audit AI decision-making processes for **accountability**	Auditability and documentation	AI systems should generate records for internal and external audits
	Protect consumers from **biased and unfair decisions** in finance	Bias detection and fairness	AI models must be tested for biases against protected groups
Internal auditors	**Verify** that AI-driven financial decisions comply with regulations	Regulatory reporting and justification	AI decisions should be easy to document for audits
	Assess AI models for **bias, fairness, and risk exposure**	Bias and discrimination monitoring	Identify and mitigate unfair biases in lending and investment decisions
		Risk mitigation and fraud detection	Validate flagged transactions or suspicious financial activities
		Benchmarking against traditional models	Compare AI-based decision making with human judgment
Risk management teams	Assess and manage **credit, market, and operational risks**	Risk transparency	Understanding the AI model's risk prediction logic
	Ensure that AI-driven financial models meet **risk assessment standards**	Scenario analysis and stress testing	Analyzing how model predictions change under different risk conditions
	Use AI-driven predictions for **portfolio risk assessment and credit underwriting**	Robustness and stability	AI models must be resilient to changes in market conditions
		Explainability for decision support	Risk teams need tools to interpret AI recommendations

(continued)

Exhibit 9. Stakeholder Needs from XAI Models in Finance *(continued)*

Stakeholder Group	Key Responsibilities	XAI Needs	Description
Data scientists and AI model developers	Build and optimize **machine learning models** for financial applications	Feature attribution and model debugging	Identifying which features influence predictions the most
	Ensure AI models are **accurate, interpretable, and efficient**	Interpretability vs. performance trade-off	Balancing accuracy with interpretability in AI models
	Develop models that align with **regulatory requirements and business needs**	Standardized XAI benchmarking	Comparing different explainability techniques
		Scalability and computational efficiency	Making explainability feasible in trading and risk management
Traders, investment analysts, and portfolio managers	Use AI models for **algorithmic trading, portfolio optimization, and risk assessment**	Explainability in trading signals, investment decisions, and recommendations	Understanding why an AI model recommends buying/selling an asset
	Identify market trends and interpret AI-generated **buy/sell recommendations**	Market trend interpretability	Explaining how macroeconomic and financial indicators influence trading decisions
		Risk exposure analysis	Understanding portfolio exposure to financial risks based on AI outputs
		Real-time decision making	Fast and interpretable AI models for high-frequency trading
Loan applicants	Understand how AI models **impact their financial decisions** (e.g., loan approvals/ rejections, credit scores)	Simple and actionable explanations	Applicants need clear feedback on why they were approved or denied
	Ensure fair treatment in decisions	Trust and fairness	Applicants need confidence that AI systems do not discriminate
	Improve approval chances	Transparency in credit decisions	Users want to know what factors they can improve to qualify for loans
Investors	Justify investment recommendations	Transparent and personalized insights	Investors want to understand why certain stocks or funds are recommended
	Manage risk exposure	Scenario analysis and risk attribution	Investors need clarity on how market shifts or inputs affect portfolio outcomes
	Build financial literacy	Simple explanations for nonexperts	Users benefit from explanations that enhance understanding of financial tools

"Creating Value from Big Data in the Investment Management Process" (Wilson 2025). Algorithmic appreciation can lead to confirmation bias, in which users accept explanations that align with their preconceived beliefs while ignoring contradictory information. To prevent misplaced trust in AI-driven insights, it is crucial to educate stakeholders on the limitations of XAI.

Another challenge is the mismatch between explanation granularity and stakeholder needs. Different users require different levels of explanation, ranging from simple cause-and-effect reasoning for customers to in-depth technical details for regulators and data scientists. Failure to provide explanations that align with stakeholder expectations can lead to misunderstandings or regulatory noncompliance.

Additionally, the lack of standardized evaluation metrics to measure the effectiveness of XAI methods further complicates implementation. For example, no universally accepted benchmark exists to measure the quality and reliability of AI-generated explanations, which leads to inconsistencies in how financial institutions assess model transparency.

Cognitive overload is another significant issue associated with XAI. If AI explanations are too complex, they can overwhelm users, making the explanations difficult to interpret and act on. Overly simplified explanations, however, may omit critical details, leading to incomplete or misleading interpretations. Financial institutions must carefully design their XAI implementations to ensure that explanations are both informative and accessible to diverse stakeholders.

Privacy risks also pose a challenge in XAI adoption. Financial AI models often rely on sensitive customer data, and providing detailed explanations of model decisions can inadvertently expose private information. For example, counterfactual explanations—statements such as "If your income were $5,000 higher, your loan would be approved"—could reveal confidential financial data. Striking a balance between transparency and data security is critical to prevent unintended privacy violations while maintaining compliance with data protection regulations.

Although such industries as health care have developed well-defined explainability guidelines, the financial sector lacks a universal standard for implementing XAI. Differences in regulatory requirements in different regions, such as the European Union and the United States, further complicate adoption. Without standardized guidelines, financial firms that operate internationally may struggle to achieve regulatory compliance and ensure fairness in AI decision making.

Crucially, financial institutions often lack user-friendly XAI tools that cater to nontechnical stakeholders. Many existing XAI models are designed primarily for data scientists rather than business professionals, regulators, or customers. In the absence of intuitive dashboards, interactive visualization tools, and natural language explanations, the accessibility of XAI insights will be limited.

Incorporating user-friendly design elements, such as plots, heatmaps, and textual summaries, can improve stakeholder engagement and comprehension.

Recommendations for Enhancing XAI in Finance

To maximize the benefits of XAI while addressing its challenges, financial institutions should adopt a strategic approach to implementation. First, the development of standardized XAI frameworks would ensure consistent high standards of practice among institutions. Collaboration between financial regulators and industry stakeholders can help establish clear guidelines for explainability.

Second, institutions should tailor AI explanations to different stakeholder groups. Customers, regulators, practitioners, and data scientists each require distinct levels of detail, and user-centric explanations can improve trust and engagement with AI models. Financial institutions should invest more widely in interactive and user-friendly XAI tools, making AI explanations more accessible to nontechnical users through dashboards, natural language summaries, and visual representations.

Finally, real-time explainability should be prioritized for AI models that influence immediate financial decisions. Financial firms can explore such techniques as local approximations and surrogate models to generate real-time explanations without compromising speed or efficiency.

Exhibit 10 provides a summary of challenges, benefits, and recommendations related to XAI.

Alternative Approaches to AI Explainability

This section explores two alternative approaches to XAI: evaluative AI and neurosymbolic AI. These approaches address the current limitations of XAI in different ways, each offering unique advantages and trade-offs.

Evaluative AI: Evidence-Based Decision Support

Proposed by Miller (2023), evaluative AI shifts from recommendation-driven decision support to a hypothesis-driven framework (see **Exhibit 11**). Instead of AI providing a single recommendation and justifying it, evaluative AI presents evidence for and against multiple options, allowing decision makers to engage in a structured evaluation process. This approach aligns with human cognitive reasoning, particularly abductive reasoning, which involves generating and testing hypotheses to make informed decisions. Evaluative AI can be particularly beneficial in high-stakes decision-making scenarios in which human oversight is critical (Miller 2023).

A key strength of the evaluative AI approach is that it reduces automation bias, in which users blindly trust AI-generated recommendations. By shifting the

Exhibit 10. Summary of XAI Benefits, Challenges, and Recommendations

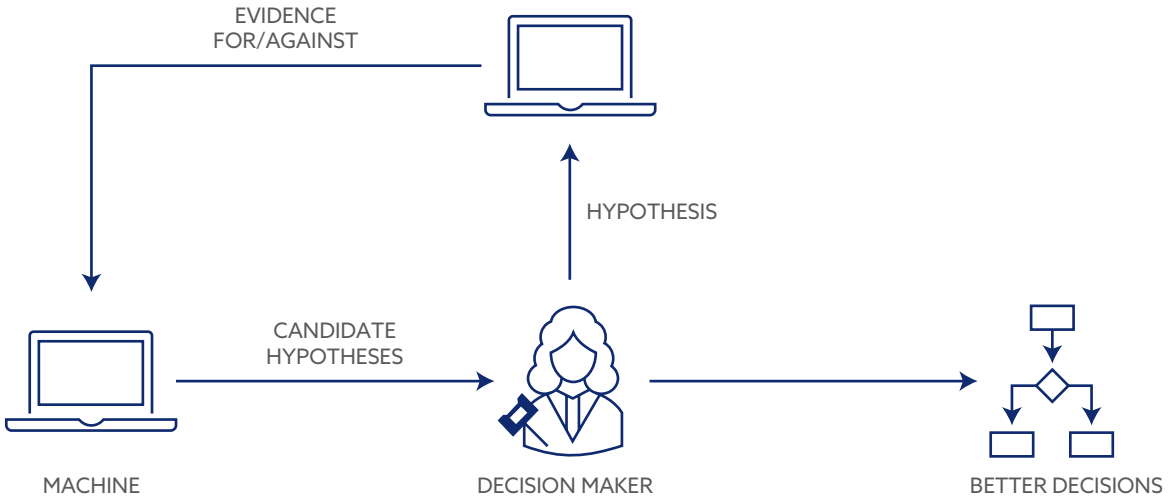
Aspect	Benefits	Challenges	Recommendations
Transparency and trust	Improve stakeholder trust through clear explanations	Users may over-rely on AI-generated explanations	Educate users on AI limitations, and encourage critical review
Model validation	Enhances model monitoring and error detection	No benchmarks for XAI explanations	Develop industry-wide benchmarks for explanation quality
Decision making	Improves risk assessments and investment decisions	Explanations may not align with user needs	Tailor explanations for different financial stakeholders
Regulatory compliance	Helps meet legal explainability requirements	Global inconsistencies in AI regulations	Align financial XAI with General Data Protection Regulation, AI Act, and local laws
Customer experience	Increases satisfaction by explaining financial decisions	Too much detail can cause cognitive overload	Use simple, interactive explanations for customers
Bias reduction	Identifies and mitigates biases in AI decisions	Some biases may remain hidden despite explanations	Conduct fairness audits on XAI models
Fraud detection	Explains fraud detection patterns effectively	Providing real-time explanations is complex	Optimize XAI for fast, interpretable fraud detection
Model performance	Enables interpretable models for financial decisions	Trade-off between accuracy and explainability	Balance performance with transparency using hybrid models
Privacy protection	Maintains security while offering model insights	Explanations risk revealing sensitive financial data	Apply privacy-preserving XAI techniques
User engagement	Helps nontechnical users understand AI models	Lack of user-friendly interfaces for XAI	Develop intuitive dashboards and NLP-based explanations

focus to user-driven hypothesis testing, this approach mitigates overreliance on AI and enhances critical thinking in decision making. It has drawbacks, however. Evaluative AI requires active engagement from users, which can lead to cognitive overload, particularly in time-sensitive situations in which users may prefer direct recommendations. Additionally, without a structured framework for presenting evidence, decision makers might still struggle to interpret and weigh the information effectively (Miller 2023). Testing of this approach has thus far been limited to the medical and maritime fields, so the extent of its efficacy for finance is not yet known.

Neurosymbolic AI: Bridging Learning and Reasoning

Neurosymbolic AI, explored by Besold, Bader, Bowman, Domingos, Garcez, Hitzler, Kühnberger et al. (2021), aims to integrate symbolic logic—a formal system for representing and reasoning with discrete, high-level concepts,

Exhibit 11. A Model of Evaluative AI



Notes: Evaluative AI explicitly provides support to explore options and perform trade-offs. In this framework, the judgment is made by the human decision maker with support from the decision-support tool, which gives feedback on (evidence for/against) proposed hypotheses.

Source: Miller (2023).

such as objects, relationships, and rules—with neural networks, enabling AI systems to both learn from data and perform logical reasoning. This approach seeks to combine the best of both worlds: the pattern recognition capabilities of deep learning and the structured reasoning of symbolic AI. By doing so, neurosymbolic AI can address the common criticism of machine learning models being opaque and lacking reasoning capabilities.

Additionally, in "Interpretable, Transparent, and Auditable Machine Learning," Philps, Tilles, and Law (2021) proposed an approach where machine learning can derive clear, human-readable investment rules. Unlike traditional factor investing, this method uses nonlinear techniques while maintaining interpretability—ensuring that AI-driven investment strategies are both effective and auditable.

One major advantage of the neurosymbolic AI approach is that it allows AI to provide rule-based justifications for its decisions. Unlike deep learning models such as neural networks, which rely on loss function optimization and back-propagation, neurosymbolic AI can explain its reasoning using structured logic, making it suitable for domains requiring strict logical consistency, such as legal reasoning, financial systems, and automated theorem proving (Besold et al. 2021). A major challenge of neurosymbolic AI, however, lies in its computational complexity. Integrating symbolic reasoning with deep learning requires significant resources, and scaling such systems to handle large datasets remains an ongoing research challenge (Besold et al. 2021).

Exhibit 12. Comparative Analysis of Alternative Approaches

Approach	Main Idea	Key Advantages	Challenges and Limitations	Example Applications
Evaluative AI (Miller 2023)	Provides evidence instead of recommendations, allowing users to evaluate hypotheses	Reduces automation bias Aligns with human cognitive reasoning Enhances critical thinking in decision making	Requires active engagement, leading to cognitive overload May still require structured evidence presentation for better usability	Medical decision support, policymaking, financial risk assessment
Neurosymbolic AI (Besold et al. 2021)	Combines symbolic reasoning with neural networks to enable AI that can both learn and reason	Allows AI to explain its reasoning using structured logic Suitable for domains requiring logical consistency Bridges learning and reasoning	Computationally expensive Difficult to scale to large datasets	Legal AI, financial systems, automated reasoning, intelligent tutoring systems

Summary

These alternative approaches to XAI offer different pathways to improving AI transparency and trustworthiness. Evaluative AI provides a user-driven, evidence-based approach, which reduces automation bias but requires more cognitive effort. Neurosymbolic AI enhances AI's reasoning capabilities, making it particularly useful for logical decision-making domains, although it faces scalability challenges.

The optimal approach depends on the application domain. For complex reasoning tasks, neurosymbolic AI provides structured explanations. Meanwhile, evaluative AI empowers human decision makers by providing evidence rather than dictating choices. Future AI development may incorporate hybrid approaches, combining the best aspects of these methodologies to create more transparent, trustworthy, and effective AI systems.

Exhibit 12 provides a comparative analysis of evaluative AI and neurosymbolic AI.

Conclusion: Moving Toward More AI Explainability

Explainable artificial intelligence is becoming increasingly crucial as AI technologies permeate critical sectors, such as health care, finance, and legal reasoning. The ability to ensure that AI models are both powerful and transparent is vital for building trust and accountability in these high-stakes domains. Although the current landscape of XAI approaches is diverse, they share a common feature: automation. Solely relying on automated solutions

may be insufficient, however, especially for nontechnical stakeholders who interact with AI systems in their daily operations. Human judgment plays a vital role in determining which XAI method is most appropriate for a given task, particularly when weighing trade-offs between interpretability, fidelity, and usability. No one-size-fits-all solution exists: Effective explainability requires thoughtful selection and contextual understanding rather than blind reliance on any single technique.

Several strategies can be implemented in order to enhance transparency and foster effective collaboration between humans and AI. Automated explanations are essential, providing users with insights into the decision-making processes of AI models. Additionally, targeted training programs can equip users with the necessary knowledge and skills to interpret and work alongside AI systems effectively. Job redesign is another crucial strategy, ensuring that human–AI collaboration is seamlessly integrated into organizational workflows.

Moreover, these strategies raise intriguing empirical research questions. For instance, what is the most effective way to support nontechnical operational users in order to improve job quality and work output quality? Focusing on work autonomy, organizational knowledge, and human learning can provide valuable insights into the impact of AI knowledge on job performance. By examining whether the presence or absence of knowledge about AI, including automated explanations, affects job quality and work output, firms can better understand the dynamics of human–AI interaction.

Ultimately, fostering a deep understanding of AI among all stakeholders will lead to more effective, transparent, and reliable deployment of AI technologies. As organizations continue to adopt AI systems, prioritizing explainability will be paramount in ensuring that these technologies are trusted and used to their full potential. By embracing a holistic approach to XAI that considers the needs of diverse stakeholder groups, we can pave the way for a future in which AI enhances human capabilities and drives innovation in high-stakes environments.

Glossary

Algorithmic transparency: This concept deals with the human user's ability to understand how the model reacts to varying inputs and, more importantly, the ability to reason about errors the model produces. Algorithmic transparency is achieved if a model can predict how changes in the input will affect the output. This is crucial for debugging and improving the model by facilitating user understanding of model behavior in different scenarios.

Decomposability: When interpretability is available in every portion of the model, including inputs, outputs, and internal parameters. A decomposable model allows a person to understand how each part of the model contributes to the final decision. For example, in a decision tree, one can examine each node and understand how it affects the overall outcome.

Example-based explanations: These provide users with specific examples of past cases that are similar to the current situation. Financial advisers and robo-advisers can use example-based explanations to justify their investment recommendations. For instance, if a system recommends a particular stock, it can provide examples of similar stocks that performed well under comparable market conditions.

Gradient boosting: An ML technique that builds a strong predictive model by combining many simple models—usually decision trees—added sequentially. Each new model focuses on correcting the errors made by the previous models, gradually improving overall accuracy. In finance, it is widely used to predict such outcomes as stock returns, credit risk, default probabilities, and fraud.

Insight portal: A centralized digital platform designed to collect, analyze, and present data-driven insights in a user-friendly way. These are typically used by businesses or organizations to support decision making, track performance, monitor trends, and share analytical findings across teams.

Reinforcement learning (RL): Inspired by how humans or agents learn by interacting with an environment and receiving feedback. An RL model tries different actions, learns from rewards or penalties, and improves over time to make better decisions—for example, developing an algorithmic trading strategy that learns over time by placing trades in a simulated market and optimizing for long-term returns while minimizing risk.

Rule-based explanations: Provide users with a set of rules that the system follows to make decisions. For example, in insurance underwriting, rule-based systems apply predefined rules based on such factors as age, health, occupation, and location to assess the risk associated with insuring an individual or a property.

Simulatability: Refers to the ability of a model to allow a human observer to simulate a thought process over its inner workings. In other words, a model is simulatable if a person can understand and mentally follow the steps the model takes to arrive at a decision. This is typically easier with simpler models, such as decision trees or linear regression, for which the decision-making process is straightforward and transparent.

Structured learning: Used when both the input and output data are organized and follow a clear structure, often with multiple related outputs that need to be predicted together. This is especially useful when the relationships between output variables matter. For example, simultaneously predicting the likelihood of loan default, the expected loss, the recovery rate, and the time to default. Because these outputs are connected, predicting them together provides more accurate and useful results.

Supervised learning: Involves training a model using historical data where the outcomes are known. The algorithm learns the relationship between input features (such as financial ratios or transaction details) and outcomes (such as "default" or "not default")—for example, predicting credit risk by training a model on past loans when you already know whether the borrower defaulted.

Unstructured learning: Typically refers to working with data that are not in a fixed format (i.e., not in neat rows and columns)—such as text, audio, images, or video—and applying algorithms that can make sense of this unstructured format. The focus is discovering meaning or structure within messy data. Examples include analyzing quarterly earnings calls or CEO interviews to identify shifts in sentiment.

Unsupervised learning: Deals with data that lack predefined labels or outcomes. The model looks for hidden patterns or groupings in the data, such as similarities, clusters, or anomalies. For example, segmenting clients based on trading activity to uncover behavioral groups, such as day traders, institutional investors, or passive long-term holders—without anyone labeling them upfront.

Acknowledgments

I thank the following individuals for sharing their knowledge and perspectives. This report, however, reflects my views and not necessarily their views or those of their employers. Any errors are my own.

I gratefully acknowledge the topic experts and thought leaders who contributed to the findings of this report, including the following:

- Tim Miller, PhD, TIET-UQ Chair of Data Science, Professor of Artificial Intelligence, University of Queensland
- Charles Wu, CFA, Chief Investment Officer, State Super
- Ellenora Webster, Quantitative Analyst, State Super
- Alexander Bakker, PhD, Senior Quantitative Analyst, State Super

I also want to express my appreciation to the following colleagues for their contributions to this report:

- James Tait, Affiliate Researcher, CFA Institute Research and Policy Center
- Peter Went, PhD, CFA, Senior Director, Learning Content, CFA Institute Learning Content

Finally, special thanks to Rhodri G. Preece, CFA, senior head of research at CFA Institute Research and Policy Center, for his meticulous review and feedback.

Appendix: Selected Case Studies

This appendix provides two interesting case studies on explainable AI implementation.

Case Study: FSF's Implementation of Explainable AI for Client Personalization in Financial Services

FSF is a global financial services firm with a reputation for both operational excellence and client-centric innovation.[4] This case study examines how FSF implemented an explainable artificial intelligence solution to personalize digital experiences for anonymous website visitors. Through the use of supervised machine learning and ensemble modeling, FSF was able to infer visitor types and deliver tailored content in real time.

Organizational Context

FSF operates globally across investment and operational domains. Within its Behavioral Marketing division—an interdisciplinary team combining quantitative analytics and client strategy—the firm identified a strategic goal: to infer the identity or user type of anonymous digital visitors (e.g., institutional investor, portfolio manager, analyst, individual investor) in order to deliver personalized content and improve client conversion rates.

Problem Statement

FSF's digital platforms attract a broad spectrum of users whose identities are often unknown at the point of contact. Understanding whether a visitor is, for example, a portfolio manager conducting due diligence or a retail investor casually browsing is critical for delivering relevant content and engagement strategies.

Traditional user segmentation approaches based on registration or user accounts were inadequate. Such approaches were tailored to an existing customer base and could not incorporate the behavioral patterns of prospective customers. Therefore, FSF sought to develop an AI solution that could infer user types based solely on anonymous behavioral data. The goal was to deliver relevant content to users in real time, enhancing engagement and conversion rates, all while maintaining compliance with stringent financial regulations.

[4]FSF is a pseudonym used throughout the case study to preserve the company's confidentiality.

Model Development

The firm adopted a *supervised machine learning* approach, training the model on historical behavioral data from known client segments. The input data included the following:

- Session length
- Number and type of pages visited
- Time spent per page
- Content themes engaged (e.g., equities, fixed income, FX)

All data were strictly anonymized and excluded any personally identifiable information. No cookies were used to track personal identity, and no sensitive information was used or stored—just behavioral signals, parsed and interpreted. Over time, a pattern emerged: The browsing habits of a known portfolio manager looked strikingly different from those of a casual investor. That insight became the cornerstone of their model.

To ensure explainability and regulatory acceptability, FSF used an *ensemble model* consisting of the following:

- *Logistic regression (logit):* selected for its simplicity and high interpretability
- *Random forest:* capable of modeling nonlinear relationships while maintaining partial transparency
- *Gradient-boosting trees:* included for enhanced predictive accuracy

Each submodel generated a classification independently. A majority-vote mechanism was then used to determine the final user type classification. This ensemble strategy balanced predictive performance with model transparency.

Model Validation

FSF used rigorous validation techniques, including the following:

- *Out-of-sample testing:* The process involved training the model on a subset of known visitors (representing 70% of the dataset) and then evaluating (or testing) it against those withheld from the training process (the remaining 30%) to assess generalization.
- *Temporal validation:* In addition, FSF applied the model to historical site traffic—scoring visitors from a year ago, then checking to see how accurate the predictions were once those individuals eventually became clients.

Not only did the model predict with a high degree of accuracy, its reasoning—rooted in observable behavior—was immediately intuitive to executives.

Deployment and Personalization Strategy

Once validated, the XAI model was integrated into FSF's digital experience platform. With leadership buy-in secured, FSF integrated the model into its live systems. Now, when a visitor lands on the site, the AI begins quietly working in the background. A highly engaged institutional investor browsing at 10:30 a.m. on a weekday might be shown detailed white papers and portfolio tools. A retail investor casually browsing late at night might be offered educational videos or fund comparisons.

This real-time personalization echoes FSF's ethos of meeting investors where they are—offering clarity, not confusion, and relevance, not noise. The approach improved engagement rates and reflected industry best practices in investor transparency and relevance.

Benefits and Challenges

As FSF rolled out its explainable AI model across its digital platforms, the Behavioral Marketing team quickly began to see tangible results—but not without a few hard-earned lessons along the way. The project demonstrated the value of transparent AI in driving both internal and client-facing outcomes, while also revealing the practical trade-offs and operational considerations that come with deploying such a system at scale. The experience highlighted the following key benefits and challenges:

Benefits:

- Improved client engagement and conversion through targeted content.
- Internal adoption and trust, driven by the transparency of model logic.
- Regulatory alignment resulting from the model's auditable structure and anonymized inputs.

Challenges:

- Model complexity versus explainability: More complex models (e.g., deep neural networks) were ruled out, despite potentially higher accuracy, because of lack of transparency.
- Behavioral drift: Visitor patterns changed over time, necessitating periodic retraining.
- Data limitations: Strict anonymization reduced available signal richness.

Conclusion

FSF's implementation of XAI for digital client engagement demonstrates how financial institutions can balance innovation, personalization, and compliance. By using an ensemble model with interpretable components, the firm

Exhibit A1. Technical Specifications of the FSF XAI Model

Feature	Description
Model type	Ensemble model
Components	Logistic regression
	Random forest
	Gradient-boosting trees
Voting mechanism	Majority vote (2 out of 3)
Data used	Anonymized behavioral tags: session time, number of pages visited, content category, etc.
Supervision type	Supervised learning (trained on known client behavior patterns)
Explainability tools	Intrinsic explainability from logit + feature importance analysis on trees
Validation approach	Out-of-sample testing
	Temporal backtesting on historical visitors
Privacy compliance	Fully anonymized dataset; no personally identifiable information used
Deployment	Real-time scoring on FSF's digital platform
Model retraining	Periodic, based on updated behavior data and business cycles

successfully predicted visitor types and delivered tailored experiences while maintaining transparency for internal stakeholders and regulators. The case underscores the importance of explainability as a prerequisite for trust in financial AI applications. **Exhibit A1** provides the technical specifications for the firm's model.

Case Study: Implementing Explainable AI in Investment Strategy at State Super

State Super is the trustee of the State Authorities Superannuation Scheme, State Superannuation Scheme, and Police Superannuation Scheme. The assets of these schemes have been combined into the STC Pooled Fund. State Super is one of Australia's oldest superannuation schemes and, as of 30 June 2024, has more than 80,000 members and $37 billion in assets. State Super's use of reinforcement learning and large language model tools for investment decision making was the subject of a case study in the CFA Institute report "Pensions in the Age of Artificial Intelligence" (Hayman 2024). Explainability was noted as a key consideration in the scheme's development of ML-driven RL models.

AI Applications at State Super

As part of its digital transformation and modernization of investment processes, the organization has developed and deployed a proprietary AI-powered platform known as the Insight Portal. This platform incorporates machine learning, including reinforcement learning techniques, to analyze structured market data and inform tactical decisions, such as country allocation tilts and currency pair selections.

The Insight Portal does not function as a replacement for legacy investment tools, such as Bloomberg or Excel. Instead, this AI system complements these platforms, offering an additional layer of analysis that surfaces relationships and forecasts that might be opaque or overlooked by human analysts. What sets the Insight Portal apart is its delivery of explanatory outputs in addition to investment recommendations—clarifying why the model suggests certain decisions. The explanatory aspect is a crucial part of the investment recommendation; it ensures sensibility in feature selection and that, importantly, decisions are not being driven by data mining.

This focus on explainability reflects a foundational principle of State Super's AI strategy: The technology must be interpretable and accountable.

Motivation for Developing an Explainable Solution

The impetus for developing an explainable AI model stemmed from both philosophical and practical considerations. As Charles Wu, the organization's chief investment officer, explained, the investment team recognized that no individual—no matter how experienced—could fully absorb and interpret the volume and complexity of today's interconnected global market data. Although human analysts naturally bring specialized expertise, they are also limited by their own perspectives and biases.

Wu noted that traditional quant models began to show limitations during periods of market anomaly, particularly during the era of negative-yielding bonds around 2017–2018.[5] These kinds of macroeconomic conditions challenged conventional valuation tools and prompted the team to seek out more adaptive, data-driven models that could offer alternative insights. Machine learning provided that opportunity—but only if its outputs could be meaningfully interpreted.

Trust and transparency were paramount. As Wu put it, "We are in the business of trust." An opaque, black-box system would not meet the needs of stakeholders such as senior leadership or be compliant with regulatory expectations. The organization needed a solution that could both uncover new patterns in the data and make those patterns intelligible to the humans responsible for managing billions in member assets.

[5]For more information on negative interest rates, see the CFA Institute Research Foundation publication *The Incredible Upside-Down Fixed-Income Market* (Bhansali 2021).

Development of the XAI Solution

State Super opted to develop the AI system entirely in-house, drawing on the capabilities of a hybrid investment–technical team. This team included two dedicated programmers and investment professionals who shared varying degrees of coding and analytical skills. According to Wu, the team's diverse yet overlapping competencies—some members being 80% technical and 20% investment, others closer to a 60/40 split—were critical to ensuring that model development was both technically robust and grounded in economic rationale.

This approach aligns closely with the CFA Institute concept of T-shaped teams (Cao 2021), which emphasizes the importance of deep expertise in a core area (the vertical bar of the "T") paired with the ability to collaborate across disciplines (the horizontal bar). By fostering team members who could bridge technical and financial domains, State Super enhanced its capacity to translate machine learning insights into economically meaningful investment decisions.

The development process began with the transformation of structured market data into usable features. This transformation required extensive feature engineering to ensure comparability across variables—P/Es, for instance, were converted into standardized scores to bring consistency to the inputs. Once processed, these features were stored to allow for efficient reuse and review.

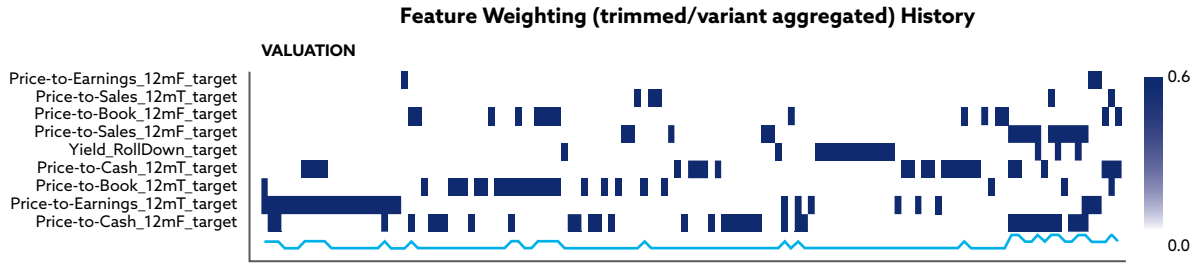
Crucially, the model was not hard-coded to replicate existing human knowledge. Instead, it was designed to seek out potentially unknown stable relationships in the data. This deliberate openness allowed the AI to offer "plays" that extended beyond traditional investment thinking, in the spirit of AlphaGo's landmark strategies in the board game Go. Yet developers remained careful not to introduce so much complexity that the model became inscrutable. The aim was to strike a balance between innovation and interpretability.

Testing the XAI Model

Once developed, the model underwent a multistage testing process that blended technical validation with investment judgment. At the coding level, developers performed several robustness and stability checks to ensure model consistency. The more critical phase of testing, however, came from the investment team, which applied domain-specific logic and judgment to evaluate the reasonableness of the model's outputs.

A key part of this testing was the use of visual analytics to promote interpretability. A fundamental design principle is that various features affect the system at different rates—some take effect rapidly, while others have a more gradual effect. **Exhibit A2** highlights how each feature has a different weight at different points in time: Each row represents the aggregate of different rates of one feature, and the *x*-axis represents time. The heatmap shows the evolution of weights assigned by the model each month (*x*-axis) to selected

Exhibit A2. Feature Weighting



Notes: Valuation indicators include forward and trailing metrics for price-to-earnings (P/E), price-to-book (P/B), price-to-sales (P/S), and price-to-cash flow (P/CF), along with a roll-down yield measure (Yield_RollDown_target). These factors are commonly used to assess relative value. Lower values typically indicate more attractively priced assets.

Yield_RollDown_target reflects the potential price appreciation of fixed income securities as they "roll down" the yield curve toward maturity, assuming an upward-sloping curve.

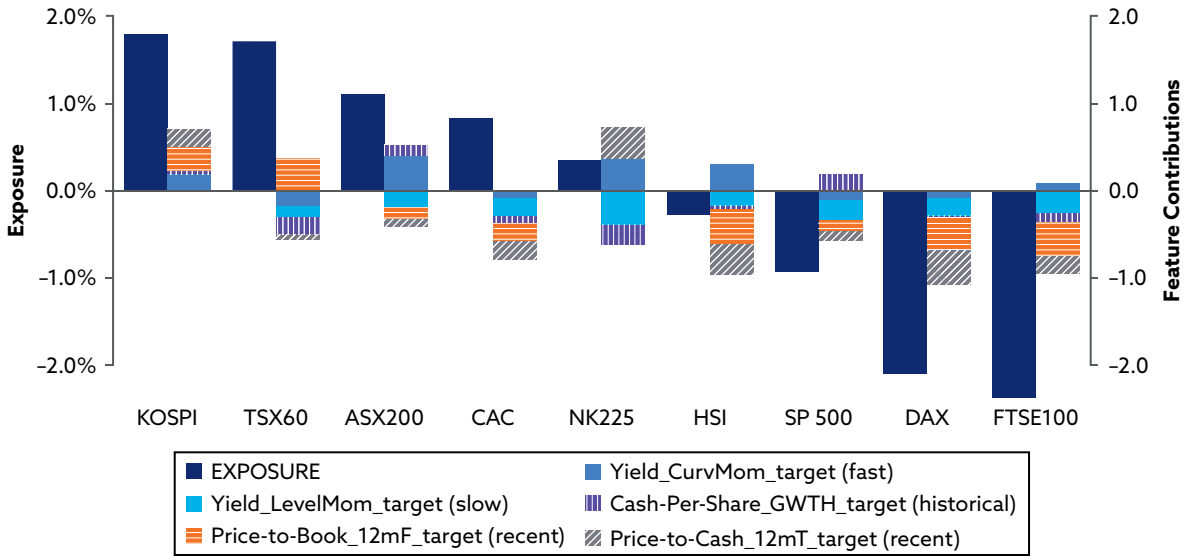
Label conventions:
12mF = 12-month forward estimate (based on forecasts)
12mT = 12-month trailing value (based on reported results)
_target = Feature used as an input to the model's predictive target

Source: State Super.

valuation features. A darker color corresponds to a higher weight. This is in line with how State Super sees the world: Sometimes certain features—such as the P/E—play a more prominent role, and sometimes their impact diminishes.

Exhibit A3 shows how different features, under trained parameters, contribute to asset exposures. In this chart, the solid blue bars represent the mean-variant

Exhibit A3. Exposure and Feature Contributions Source: State Super, 14 April 2025



Note: See Exhibit A4 for more details on variable names.

Source: State Super.

exposure for each equity index (e.g., KOSPI, TSX60, DAX). Mean-variant exposure is the percentage allocation of each financial asset—essentially, the model's reaction to available information for that asset. The direction and size of the blue bars indicate what markets the model favors (overweight exposure) or avoids (underweight or short exposure).

The stacked colored bars are designed to highlight the individual feature contributions to the model reaction; they are a joint representation of both the underlying feature data value (or "score" of said feature) and the trained weight associated with that feature. Thus, the objective of these bars is to demonstrate the primary drivers of the model reactions. Exhibit A3 shows that Korea Composite Stock Price Index (KOSPI) is a preferred exposure.

Exhibit A4 summarizes the various factor drivers covering valuation, macro momentum, and fundamental growth dimensions.

Inconsistencies in the model's behavior—especially those that lacked a plausible economic explanation—triggered review discussions. Finally, the model undergoes what Wu describes as the "sanity test" to ensure that results make intuitive and analytical sense.

Exhibit A4. Factor Comparison Table

Legend Label	Description	Time Horizon	Factor Type	Typical Use/Signal
Price-to-Book_12mF_ target (recent)	Price/projected book value using forward 12-month analyst estimates	Forward (12 months)	Valuation (forward)	Identifies undervalued stocks based on future fundamentals; lower values = more attractive
Price-to-Cash_12mT_ target (recent)	Price/trailing 12-month realized cash flow	Backward (12-month trailing)	Valuation (trailing)	Captures operational cash efficiency; lower values often signal higher quality or value
Yield_CurvMom_ target (fast)	Slope change in the yield curve (e.g., 10Y–2Y) over recent months	Short term (1–3 months)	Macro momentum (fast)	Reflects sensitivity to tactical monetary policy shifts and curve steepening/flattening
Yield_LevelMom_ target (slow)	Trend in the absolute level of interest rates (e.g., 10-year yield rising/falling)	Long term (6–12+ months)	Macro momentum (slow)	Measures exposure to interest rate regimes; affects sector or style tilts (growth/value)
Cash-Per-Share_ GWTH_target (historical)	Growth in cash per share over a historical period	Historical	Fundamental growth	Indicates strong profitability and reinvestment ability; higher growth = higher quality

Source: CFA Institute.

Implementation and Use Across the Organization

Once validated, the model is made accessible across the investment division. Although the platform is technically available to all employees at State Super, it is primarily used by portfolio analysts and investment managers, with Wu as the primary C-suite user and internal champion. The implementation was guided by a formal model policy that delineated roles, permissions, and processes for promoting models from development to final stage.

Wu emphasized that having an internal sponsor was critical for successful implementation. Because he served as both the key user and advocate for the system, the AI initiative had top-down support from the outset. This approach helped ensure organizational buy-in and minimized resistance from other teams.

The use of the Insight Portal is framed as an augmentative rather than replacement tool—one that provides an additional lens for analyzing global markets, especially in unfamiliar or complex environments. As the team becomes more comfortable with the system's foundational capabilities, more advanced models and techniques are gradually introduced.

Results

The transition to an explainable AI framework has yielded considerable benefits for State Super. Chief among them is enhanced trust—both within the investment team and across the broader organization. Because the system is transparent and its outputs can be explained, it avoids the reputational and operational risks associated with opaque "black-box" technologies. The model also serves as a learning tool, helping analysts challenge their own assumptions and uncover new insights in the data.

The path has not been without challenges, however. Prioritizing explainability has required the team to forgo some potentially more accurate but less interpretable algorithms. Another challenge has been managing expectations. In contrast to the impulse to take "big leaps" with solutions, State Super has adopted a deliberately incremental approach. Each new layer of complexity is introduced only after the foundational model has been fully understood and validated.

References

Besold, Tarek R., Sebastian Bader, Howard Bowman, Pedro Domingos, Artur d'Avila Garcez, Pascal Hitzler, Kai-Uwe Kühnberger, et al. 2021. "Neural-Symbolic Learning and Reasoning: A Survey and Interpretation." In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, edited by P. Hitzler and M. K. Sarker, 1–51. IOS Press. doi:10.3233/FAIA210348.

Bhansali, Vineer. 2021. *The Incredible Upside-Down Fixed-Income Market: Negative Interest Rates and Their Implications*. Charlottesville, VA: CFA Institute Research Foundation. https://rpc.cfainstitute.org/research/foundation/2021/negative-interest-rates.

Biddle, Justin B. 2016. "Inductive Risk, Epistemic Risk, and Overdiagnosis of Disease." *Perspectives on Science* 24 (2): 192–205. doi: https://doi.org/10.1162/POSC_a_00200.

Brennen, Andrea. 2020. "What Do People Really Want When They Say They Want 'Explainable AI?' We Asked 60 Stakeholders." In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*, 1–7. New York: Association for Computing Machinery. doi:10.1145/3334480.3383047.

Cai, Fang, and Sharjil Haque. 2024. "Private Credit: Characteristics and Risks." FEDS Notes, Board of Governors of the Federal Reserve System (23 February). doi:10.17016/2380-7172.3462.

Cao, Larry. 2021. "T-Shaped Teams: Organizing to Adopt AI and Big Data at Investment Firms." CFA Institute (30 August). https://rpc.cfainstitute.org/research/reports/t-shaped-teams.

Cao, Longbing. 2023. "AI in Finance: Challenges, Techniques, and Opportunities." *ACM Computing Surveys* 55 (3): 1–38. doi:10.1145/3502289.

Černevičienė, Jurgita, and Audrius Kabašinskas. 2024. "Explainable Artificial Intelligence (XAI) in Finance: A Systematic Literature Review." *Artificial Intelligence Review* 57. doi:10.1007/s10462-024-10854-8.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–94. New York: Association for Computing Machinery. doi:10.1145/2939672.2939785.

Elton, Daniel C. 2020. "Self-Explaining AI as an Alternative to Interpretable AI." In *Artificial General Intelligence: 13th International Conference, AGI 2020, St. Petersburg, Russia, September 16–19, 2020, Proceedings*, edited by B. Goertzel, A. Panov, A. Potapov, and R. Yampolskiy, 95–106. Cham, Switzerland: Springer. doi:10.1007/978-3-030-52152-3_10.

Fearn, Nicholas. 2024. "Can AI Outperform a Wealth Manager at Picking Investments?" *Financial Times*, 2 July. https://www.ft.com/content/3b443015-25e1-4a13-b68f-ec769934ec75.

Gerlings, Julie, Arisa Shollo, and Ioanna Constantiou. 2021. "Reviewing the Need for Explainable Artificial Intelligence (xAI)." In *Proceedings of the 54th Hawaii International Conference on System Sciences*, 1284–91 (5 January). doi:10.24251/HICSS.2021.156.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. "Explaining Explanations: An Overview of Interpretability of Machine Learning." In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. doi:10.1109/DSAA.2018.00018.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51 (5): 1–42. doi:10.1145/3236009.

Gunning, David, and David W. Aha. 2019. "DARPA's Explainable Artificial Intelligence Program." *AI Magazine* 40 (2): 44–58. doi:10.1609/aimag.v40i2.2850.

Hayman, Genevieve. 2024. "Pensions in the Age of Artificial Intelligence." CFA Institute (17 December). https://rpc.cfainstitute.org/research/reports/2024/pensions-in-the-age-of-ai.

Hoffman, Robert R., Gary Klein, and Shane T. Mueller. 2018. "Explaining Explanation for 'Explainable AI.'" *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62 (1): 197–201. doi:10.1177/1541931218621047.

Hong, Sungsoo Ray, Jessica Hullman, and Enrico Bertini. 2020. "Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW1): 1–26. https://doi.org/10.1145/3392878.

Information Commissioner's Office and The Alan Turing Institute. 2020. "Explaining Decisions Made with AI." https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/.

Joseph, John, William Ocasio, and Mary-Hunter McDonnell. 2014. "The Structural Elaboration of Board Independence: Executive Power, Institutional Logics, and the Adoption of CEO-Only Board Structures in US Corporate Governance." *Academy of Management Journal* 57 (6): 1834–1858. https://www.jstor.org/stable/43589332.

Karapiperis, Dimitris DeFrain. 2019. "Intelligent Machines and the Transformation of Insurance." *CIPR Newsletter* (January): 11–16.

Krishna, Satyapriya, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2022. "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective." arXiv (2 February). https://arxiv.org/pdf/2202.01602.

Kuiper, Ouren, Martin van den Berg, Joost van der Burgt, and Stefan Leijnen. 2022. "Exploring Explainable AI in the Financial Sector: Perspectives of Banks and Supervisory Authorities." In *Benelux Conference on Artificial Intelligence*, 105–119. Cham: Springer International Publishing. https://arxiv.org/pdf/2111.02244.

Le, Phuong-Hang Quynh, Meike Nauta, Van Bach Nguyen, Shreyasi Pathak, Jörg Schlötterer, and Christin Seifert. 2023. "Benchmarking eXplainable AI: A Survey on Available Toolkits and Open Challenges." *Journal of AI Research and Applications* 15 (2): 150–75.

Liesenfeld, Andreas, and Mark Dingemanse. 2024. "Rethinking Open Source Generative AI: Open-Washing and the EU AI Act." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1774–1787. https://dl.acm.org/doi/pdf/10.1145/3630106.3659005?trk=public_post_comment-text.

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery." *Queue* 16 (3): 31–57. doi:10.1145/3236386.3241340.

Miller, Tim. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267 (February): 1–38. doi:10.1016/j.artint.2018.07.007.

Miller, Tim. 2023. "Explainable AI Is Dead! Long Live Explainable AI! Hypothesis-Driven Decision Support." In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. doi:10.48550/arXiv.2302.12389.

Miller, Tim, Piers Howe, and Liz Sonenberg. 2017. "Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences." arXiv (5 December). doi:10.48550/arXiv.1712.00547.

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. 2018. "Methods for Interpreting and Understanding Deep Neural Networks." *Digital Signal Processing* 73 (February): 1–15. doi:10.1016/j.dsp.2017.10.011.

Moreno, Verónica. 2024. "Is XAI the Right Explanation? Concerns on Using XAI for Legally Required Explanations on Opaque Algorithmic Decisionmaking." In *Proceedings of 4th Workshop on Explainable AI in Finance (XAIFIN-2024)*.

Mueller, Shane T., Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. "Explanation in Human-AI Systems: A Literature Meta-Review." arXiv (5 February). doi:10.48550/arXiv.1902.01876.

OECD. 2023. "Generative Artificial Intelligence in Finance." OECD Artificial Intelligence Papers, No. 9, OECD Publishing, Paris. https://doi.org/10.1787/ac7149cc-en.

Olson, Lincoln. 2025. "The 9 Largest Private Credit Funds in the World (Top Firms by AUM)." Stock Analysis (4 February). https://stockanalysis.com/article/largest-private-credit-funds/.

Páez, Andrés. 2019. "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)." *Minds and Machines* 29 (3): 441–59. doi:10.1007/s11023-019-09502-w.

Philps, Daniel, David Tilles, and Timothy Law. 2021. "Interpretable, Transparent, and Auditable Machine Learning: An Alternative to Factor Investing." *Journal of Financial Data Science* 3 (4): 84–100. doi:10.3905/jfds.2021.1.077.

Preece, Alun. 2018. "Asking 'Why' in AI: Explainability of Intelligent Systems—Perspectives and Challenges." *Intelligent Systems in Accounting, Finance & Management* 25 (2): 63–72. doi:10.1002/isaf.1422.

Preece, Alun, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. "Stakeholders in Explainable AI." arXiv (29 September). doi:10.48550/arXiv.1810.00184.

Preece, Rhodri G. 2022. "Ethics and Artificial Intelligence in Investment Management: A Framework for Professionals." CFA Institute (14 October). https://rpc.cfainstitute.org/research/reports/2022/ethics-and-artificial-intelligence-in-investment-management-a-framework-for-professionals. doi:10.56227/22.1.15.

Ribera, Mireia, and Agata Lapedriza. 2019. "Can We Do Better Explanations? A Proposal of User-Centered Explainable AI." In *Joint Proceedings of ACM IUI 2019 Workshops*. https://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 1135–44. New York: Association for Computing Machinery. doi:10.1145/2939672.2939778.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." arXiv (22 September). doi:10.48550/arXiv.1811.10154.

Sudjianto, Agus, and Aijun Zhang. 2021. "Designing Inherently Interpretable Machine Learning Models." arXiv (2 November). https://arxiv.org/abs/2111.01743

Taylor, Isaac. 2024. "New AI Technology Spurs Excitement and Concerns Among Private-Credit Managers." *Wall Street Journal* (28 June). https://www.wsj.com/articles/new-ai-technology-spurs-excitement-and-concerns-among-private-credit-managers-f3a2e65f.

Tomsett, Richard, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. "Interpretable to Whom? A Role-Based Model for Analyzing Interpretable Machine Learning Systems." arXiv (20 June). doi:10.48550/arXiv.1806.07552.

van den Berg, Martin, and Ouren Kuiper. 2020. "XAI in the Financial Sector: A Conceptual Framework for Explainable AI (XAI)." Hogeschool Utrecht, Lectoraat Artificial Intelligence (September). www.researchgate.net/publication/344079379_XAI_in_the_Financial_Sector_A_Conceptual_Framework_for_Explainable_AI_XAI.

van der Waa, Jasper, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. "Evaluating XAI: A Comparison of Rule-Based and Example-Based Explanations." *Artificial Intelligence* 291: Article 103404. https://doi.org/10.1016/j.artint.2020.103404.

Wanner, Jonas, Lukas-Valentin Herm, and Christian Janiesch. 2020. "How Much Is the Black Box? The Value of Explainability in Machine Learning Models." In *Proceedings of the Twenty-Eighth European Conference on Information Systems (ECIS2020)—A Virtual AIS Conference*. https://aisel.aisnet.org/ecis2020_rip/85.

Weber, Patrick, K. Valerie Carl, and Oliver Hinz. 2024. "Applications of Explainable Artificial Intelligence in Finance—A Systematic Review of Finance, Information Systems, and Computer Science Literature." *Management Review Quarterly* 74 (June): 867–907. doi:10.1007/s11301-023-00320-0.

White House OSTP (Office of Science and Technology Policy). 2022. *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*. Washington, D.C.: Executive Office of the President of the United States (October). Archived at Internet Archive. https://web.archive.org/web/20230201095406/https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

Wilson, Cheryll-Ann. 2025. "Creating Value from Big Data in the Investment Management Process: A Workflow Analysis." CFA Institute (13 January). https://rpc.cfainstitute.org/research/reports/2025/creating-value-from-big-data-in-the-investment-management-process. doi:10.56227/25.1.7.

Wilson Drakes, Cheryll-Ann. 2021. "Explainable AI for Non-Experts: Is This a Chimera?" *Academy of Management Proceedings* 2021 (1). doi:10.5465/AMBPP.2021.14234abstract.

Yeo, Wei Jie, Wihan Van Der Heever, Rui Mao, Erik Cambria, Ranjan Satapathy, and Gianmarco Mengaldo. 2025. "A Comprehensive Review on Financial Explainable AI." *Artificial Intelligence Review* 58, Article No. 189. doi:10.1007/s10462-024-11077-7.

Author

Cheryll-Ann Wilson, PhD, CFA
Senior Affiliate Researcher
CFA Institute

ABOUT THE RESEARCH AND POLICY CENTER

CFA Institute Research and Policy Center brings together CFA Institute expertise along with a diverse, cross-disciplinary community of subject matter experts working collaboratively to address complex problems. It is informed by the perspective of practitioners and the convening power, impartiality, and credibility of CFA Institute, whose mission is to lead the investment profession globally by promoting the highest standards of ethics, education, and professional excellence for the ultimate benefit of society. For more information, visit https://rpc.cfainstitute.org/en/.

First page photo credit: Getty Images/Ryzhi

CFA Institute

PROFESSIONAL LEARNING QUALIFIED ACTIVITY

This publication qualifies for 1.25 PL credits under the guidelines of the CFA Institute Professional Learning Program.