Est. 1841 | YORK ST JOHN UNIVERSITY

Pimentel, Rita and Pisoni, Galena (2025) Rethinking explainable AI in financial services. AI & Society.

Downloaded from: https://ray.yorksj.ac.uk/id/eprint/11920/

The version presented here may differ from the published version or version of record. If you intend to cite from the work you are advised to consult the publisher's version: https://doi.org/10.1007/s00146-025-02315-9

Research at York St John (RaY) is an institutional repository. It supports the principles of open access by making the research outputs of the University available in digital form. Copyright of the items stored in RaY reside with the authors and/or other copyright owners. Users may access full text items free of charge, and may download a copy for private study or non-commercial research. For further reuse terms, see licence terms governing individual outputs. Institutional Repositories Policy Statement

RaY

Research at the University of York St John

For more information please contact RaY at ray@yorksj.ac.uk

Rethinking Explainable AI in Financial Services

The financial institutions have increasingly relied on data-driven decision-making, and in recent years, artificial intelligence (AI) has become an indispensable tool in analyzing complex financial patterns, leveraging its capabilities to process large volumes of structured and unstructured data. This increased reliance on AI has been driven by the digitization of financial products such as credit cards, loan applications, and online banking. However, its growing influence has also sparked concerns over transparency, accountability, and compliance. Explainable AI (XAI) – methods and techniques that aim on making AI models more transparent and understandable to humans – has emerged as a potential solution, promising to open the AI 'black box' and provide insights into decision-making processes. But does XAI truly deliver on this promise?

XAI includes two branches. On the one hand, there are inherently interpretable models which are designed with transparency in mind, allowing users to directly understand their decision-making process without the need for additional explanation techniques. On the other hand, there are post-hoc explainability methods that provide insights into the inner workings of complex AI models after they have been trained, helping to interpret their predictions without altering their underlying structure. Although XAI encompasses both segments, it is more commonly associated with the latter, especially among the general public.

Weber et al. (2024) in their recent survey in XAI in Finance and our own reading of several papers in the area reveal that the adoption of AI in finance has presented numerous challenges, including discriminatory biases in data outputs and regulatory expectations. To some extent, the application of XAI can improve users' understanding of the outputs generated by the AI system, allowing for improved strategies in customer relationships, market size, and fraud detection. However, XAI faces limitations in achieving strong results in certain financial areas, such as credit risk and follower investment, due to the complexity of investigating their features.

Several issues have been brought up since XAI's inception, such as causality - explanations based on statistical attributions rather than causal reasoning; replicability - explanations derived from certain algorithms can vary, making it difficult to replicate results consistently; and low-quality information - some explanations may be erroneous, misleading, or incomplete. If financial institutions use AI models that can only be partially explained, how can they justify their decisions to regulators, auditors, or affected customers? This raises the question of whether XAI actually solves the black-box problem or simply repackages it in a different form.

The financial sector operates under significant regulatory scrutiny, demanding transparency in decision-making. But explainability must serve all end-users – whether they are investors, customers, or regulators. Many XAI techniques are developed with data scientists and AI engineers in mind rather than the people impacted by financial decisions. Their outputs may make sense to a developer but not to a loan applicant denied credit or an investor questioning an automated trading strategy. As Hadji Misheva et al. (2021) show, XAI has been primarily used by AI developers as a debugging tool to check for coding issues rather than as a genuine mechanism for ensuring fairness and interpretability. This narrow use case highlights a major shortcoming: XAI's role is often confined to internal model verification rather than providing meaningful transparency to regulators and end-users. The process of debugging may help reduce biases, but it does not necessarily prevent discriminatory or opaque decision-making

practices from emerging. The need for user-centered explanations is particularly pressing in credit scoring, where decisions directly impact individuals' financial opportunities. If a rejected applicant is presented with a vague or overly technical justification, the intended transparency is meaningless. Effective XAI must move beyond technical explainability to provide intuitive, actionable insights for end-users. This remains an unresolved challenge, as most current XAI approaches fail to bridge this gap.

Regulatory bodies worldwide are implementing new laws to prevent AI applications from causing harm to individuals, both on the local and global level. The European Union has been proactive with regulations like the General Data Protection Regulation (GDPR) and the AI Act, which emphasize fairness, transparency, and accountability in AI-driven decision-making. Thus, financial institutions deploying AI models, besides of adhering to their several specific regulations, also need to comply with these general AI-related regulations. This places an additional pressure on financial institutions to adopt explainable and accountable AI systems. However, the lack of consensus on what constitutes a 'sufficient' explanation under these regulations continues to create challenges. Regulators require more than just partial, statistical justifications – they demand clear, reproducible, and legally sound explanations. The current state of post-hoc XAI may not guarantee compliance, leaving financial firms vulnerable to legal scrutiny.

The road ahead: from explainability to accountability

We recognize that XAI is an important addition to explain AI models. However, XAI in finance must be more than a technical exercise – it should consider real-world usability and how to integrate accountability. The current landscape suggests that while post-hoc explainability may offer insights, it does not truly resolve the black-box dilemma yet.

A key factor in improving the effectiveness of XAI in finance is ensuring that AI models are developed in collaboration with domain experts. Financial professionals, economists, and regulatory specialists must work alongside AI developers to create systems that not only function optimally but also align with industry needs and compliance requirements. Without this collaboration, AI models risk being detached from the realities of financial decision-making, potentially leading to flawed interpretations and regulatory misalignment.

There is eventually the need to develop XAI models that serve better the end-users and can be understood by them. Here, a close collaboration with social science as suggested by Tim (2019) may help on bringing light to what is a useful explanation. Also, involving a kind of 'AI translator' – a specialist who can bridge the gap between AI and XAI developers and financial end-users by converting complex XAI outputs into accessible and meaningful insights – could add value to existing XAI models.

A provocative alternative may be to shift from post-hoc XAI to inherently interpretable AI models that integrate transparency by design. While in some cases the better performance of highly complex models is undoubted, on others it is comparable to simpler and understandable models. In these cases, rather than using complex models and compensate for their opacity by providing explanations after the fact, financial institutions could prioritize models that are explainable from the outset.

For the financial industry to move beyond the illusion of AI explainability and focus on true accountability, concrete steps must be taken, including regulatory alignment, stronger interdisciplinary collaboration, and prioritizing inherently interpretable models.

References:

Hadji Misheva, B., Jaggi, D., Posth, J.-A., Gramespacher, T., & Osterrieder, J. (2021). Audience-dependent explanations for ai-based risk management tools: A survey. Frontiers in Artificial Intelligence, (p. 186).

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38.

Weber, Patrick, K. Valerie Carl, and Oliver Hinz. "Applications of explainable artificial intelligence in finance – a systematic review of finance, information systems, and computer science literature." *Management Review Quarterly* 74.2 (2024): 867-907.