

# Position Paper: Explainable AI (XAI) in Financial Crimes Detection

GUANMING SHI, Max Planck Institute for Security and Privacy, Germany

YIXIN ZOU, Max Planck Institute for Security and Privacy, Germany

Online financial transactions have lowered barriers for criminals to propagate and conceal financial crimes such as fraud and money laundering. In response, technology companies, financial institutions and governments have invested heavily in AI-driven systems that detect and investigate malicious transactions. Despite these investments, financial crime detection systems continue to suffer from high false-positive rates and limited forms of explainability that fall short of stakeholders' operational needs, providing insufficient support for the humans who rely on these systems to make decisions. This position paper argues that explainability in financial crime detection should be treated as a stakeholder-sensitive design challenge rather than an add-on feature. We contend that effective XAI must support stakeholders in understanding how a system reaches its decisions in ways that align with their operational workflows. Finally, we propose future research directions for designing XAI that meaningfully support stakeholders engaged in or affected by high-stakes decision-making in financial crime detection.

CCS Concepts: • **Human-centered computing** → **Explainable AI**.

Additional Key Words and Phrases: Financial Crime, Explainable AI, Stakeholder Desiderata, Presented at the Human-centered Explainable AI Workshop (HCXAI) @ CHI 2026

## 1 Introduction

Online financial transactions underpin many contemporary technological advances, including e-commerce and digital payment systems, and have become deeply embedded in everyday economic activity. While these systems offer speed and convenience, their widespread adoption has also created opportunities for criminals to perpetrate financial crime at scale. This is because the velocity, volume, and cross-border nature of online transactions enable illicit financial flows to be concealed across complex digital infrastructures.

The International Monetary Fund (IMF) [21] describes financial crime as an illegal activity that damages financial systems and exploits regulatory frameworks. The most common types of financial crimes are fraud, money laundering, tax evasion, and insider trading [3]. Global losses attributable to fraud and money laundering amount to billions of dollars annually, e.g., USD 21 billion in the United States [12] and €7.4 billion in the European Union [19] just in 2024. Beyond direct monetary losses, financial crimes fuel broader societal harms, including organized crime and drug trafficking [48]. Monetary and psychological costs of financial crimes are often borne by companies and private individuals, many of whom are unable to fully recover their losses. Even when reimbursement is provided, victims may face significant administrative burdens, reputation losses, and psychological stress following victimization [4]. Together, these impacts underscore financial crimes as a persistent and high-stakes societal challenge.

In response, governments, financial institutions, and technology companies have turned to Artificial Intelligence (AI) systems to support the detection of online financial crimes. These systems offer powerful capabilities for identifying anomalous transaction patterns and predicting risky behavior. However, their practical deployment is often constrained by high false-positive rates and limited transparency. Wedge et al. [46] quantitatively demonstrated that one in five blocked financial transactions was fraudulent and that every sixth transaction was mistakenly flagged. When system outputs cannot be meaningfully interpreted, particularly in the presence of frequent errors, trust in automated decisions erodes, and human operators struggle to act on model recommendations.

---

Authors' Contact Information: Guanming Shi, [guanming.shi@mpi-sp.org](mailto:guanming.shi@mpi-sp.org), Max Planck Institute for Security and Privacy, Bochum, Germany; Yixin Zou, Max Planck Institute for Security and Privacy, Bochum, Germany, [yixin.zou@mpi-sp.org](mailto:yixin.zou@mpi-sp.org).

Table 1. Classes of stakeholders and their desiderata in applying XAI to financial crime detection, adapted from Arrieta et al. [7] and Langer et al. [27].

| Stakeholder Groups | Financial Crimes Detection Context   | Desideratum    | Prior Work          |
|--------------------|--|----------------|---------------------|
| Analysts           | Frontline anti-fraud and anti-money laundering analysts responsible for reviewing alerts and making decisions. | Trust          | [1] [8] [18] [35]   |
|                    |  | Accuracy       | [5] [9] [24] [25]   |
|                    |  | Speed          | [20] [42] [44] [45] |
| Affected Parties   | Individuals and organizations whose transactions are blocked by detection systems.                             | Fairness       | [1] [7] [16] [24]   |
|                    |  | Understanding  | [5] [17] [38]       |
| Regulators         | Financial regulators overseeing compliance.  | Accountability | [32] [36] [41]      |
|                    |  | Transparency   | [39]                |

This challenge has brought renewed attention to explainable artificial intelligence (XAI) in financial crime detection. Yet, we argue that existing approaches in this space often prioritize technical transparency over the operational and cognitive needs of the diverse stakeholders who rely on these systems. As a result, explanations risk overwhelming users without supporting effective judgment or action in practice.

## 2 Stakeholders involved in Financial Crimes Detection

Prior work has established that a lack of consensus around the meaning of explainability stems from the existence of distinct stakeholder communities, each with its own intentions and requirements [40]. Building on this, Arrieta et al. [7] propose a taxonomy that distinguishes between users, system developers, and affected parties. While these groups may share overlapping goals, such as understanding system behavior or ensuring regulatory compliance, their needs for explainability mostly differ. Langer et al. [27] conceptualize the goal, expectations, and demands of different classes of stakeholders as “stakeholders’ desiderata” and argue that the primary objective of explainability approaches is to satisfy stakeholder desiderata. Drawing from their works [7, 27], we identify and distinguish three classes of stakeholders and their corresponding desiderata in Table 1: *analysts* who review and act on system outputs; *regulators* who oversee compliance and accountability and *affected parties* whose transactions or activities are subject to those decisions.

### 2.1 Explanations Do Not Fit Users in High-Stakes Environment

In financial crimes detection, members of the “analysts” stakeholder class are the anti-fraud or anti-money laundering officers within law enforcement, technology companies, or banks. Examples of users include officers in INTERPOL’s Financial Crime and Anti-Corruption Center, Singapore Police Force’s Anti-Scams Center, and Trust and Safety teams in Google, Meta, and ByteDance. These users take into account recommendations of AI to make decisions, yet they are usually not experts regarding the technical details and the datasets that were used to train their systems. Despite this, they are held accountable for reviewing and acting upon model outputs, often under time pressure and regulatory scrutiny. In addition, failures in detection can lead to wrongful arrests, unjustified account freezes and a broader erosion of trust in their institutions. As a result, trust [8, 18], accuracy [25], and speed [20] emerge as central desiderata for explainability among this stakeholder group.

Current systems attempt to address this need through post-hoc explainability techniques such as LIME [43], SHAP [30], or explanations generated via Large Language Models (LLMs) [47]. However, prior work suggests that these approaches often fail to align with users’ operational realities. Explanations that are lengthy, overly technical, or detached from investigative workflows risk overwhelming analysts rather than supporting them [17, 28, 35]. In other

similar high-stakes domains, such as cyber security [42] and healthcare [44], explanations generated using LLMs have been shown to be ineffective in practice, as teams often lack the time required to read and interpret verbose outputs [31, 45]. Such misaligned explanations do not support the core desiderata of trust, accuracy, and speed, but instead contribute to cognitive burden and reduced trust in AI-assisted decision-making.

## **2.2 Limited Transparency for Affected Parties**

Affected parties in financial crime detection include individuals and organizations whose transactions are monitored, flagged, or blocked by automated systems. These stakeholders do not have direct interaction with detection systems and have limited insight into how decisions are made. Yet, they experience the consequences of these decisions most directly, often without prior warning or meaningful opportunity for recourse. As such, they are most concerned with the fairness of decisions made, whether they were treated consistently and without bias.

There has been little research done into the types of explanations provided to affected parties in financial crime contexts. Due to security and legal constraints [13], explanations are deliberately vague and frequently framed in procedural terms that do not support lay understanding. As a result, affected parties are left unable to assess whether a decision was justified or erroneous, undermining perceptions of fairness. The resulting opacity has been found to be a barrier to participation in digital payment methods in movements to a cashless society [14]. Financial regulators play a key role in establishing standards that require banks to make their decisions transparent, ideally in consultation with end users to ensure that explanations meet their needs. The Payment Services Directive 2 (PSD2) reflects a step in this direction, requiring payment service providers in the EU to notify users and provide information about actions affecting financial transactions. However, PSD2 does not define a specific format or level of detail for these explanations, leaving providers with substantial discretion in how they communicate with customers. To further improve on these initiatives, academia and government can work together to design human-centric explanations evaluated by end users to ensure that they are understandable through accessible language.

## **2.3 Lost Trail of Accountability for Regulators**

Regulators are responsible for overseeing the deployment and use of AI systems in financial crime detection, with mandates around legal and regulatory compliance. Unlike frontline users, regulators are less concerned with the correctness of individual transaction decisions and more focused on whether detection processes as a whole operate within legal, accountability, and privacy boundaries. Use Singapore as an example, this role is exemplified by the Monetary Authority of Singapore, which supervises the use of AI-enabled systems by financial institutions. For this stakeholder group, the primary desideratum is accountability, as regulators must be able to determine where responsibility lies when harm or error occurs. With increasing use of artificial systems, accountability gaps might emerge [32, 41]. For example, an analyst acting on the outputs of a fraud detection system may not know that the output was wrong, so blaming them for ensuing problems would ignore the AI system's contribution to the problem. Opaque systems will only amplify this issue. Hence, explanations that highlight the AI system's decision-making trail will help regulators apply legislation to the parties responsible.

### 3 Future Research Directions for XAI in Financial Fraud Detection

#### 3.1 Understanding Users’ Needs for Explanations

One problem concerning evaluating XAI methods is the insufficient consideration of human factors. Current evaluation approaches analyze only specific properties of XAI methods, without accounting for their interactions with stakeholders. For example, Keane et al.’s review shows that only 36 out of 127 research works employing counterfactual explanations adopted a human evaluation approach [23]. Even in existing user studies, users are typically “passive recipients” of explanations, as they are expected to judge more on the comprehensibility of explanations in a hypothetical scenario, instead of using it in problem-solving in their professional workflows. While some studies have evaluated the impact of AI system explanations on humans compared to scenarios where no explanations were provided [15, 29, 33], there is a need for more work on the topic, especially in time-constrained, high-stakes domains like fraud detection.

We are planning and conducting an interview study that seeks to deepen the understanding of real-world conditions for applying XAI to financial crime detection. At the time of writing, recruitment for this study has already begun. Through semi-structured interviews with practitioners involved in financial crime detection in technology companies and law enforcement, we aim to explore how AI-generated explanations are currently incorporated into investigative workflows. In particular, we seek to identify what types of information users find useful or burdensome, how explanations interact with time pressure and accountability constraints, and where mismatches arise between existing XAI techniques and practitioners’ operational needs. Insights from this study will be used to inform the design of stakeholder-sensitive XAI approaches that better support decision-making in high-stakes environments.

#### 3.2 Creating human-centric explanations

Although a wide range of XAI techniques have been introduced, they often produce explanations that are technically faithful but cognitively misaligned with how laypeople reason about decisions. Prior work has shown that explanations are most effective when they are selective, contrastive, and socially grounded, reflecting how humans naturally reason about causes and decisions rather than exposing raw model internals [35]. However, many existing XAI methods in financial domains still rely on feature attributions or statistical summaries that presuppose technical expertise and fail to extrapolate beyond the model’s input space. Such explanations place the burden of interpretation on users, who must construct their own narratives under time pressure and uncertainty.

A direction for HCXAI is the development of explanation approaches that operate at the level of concepts rather than individual features. In financial crime detection, this could involve explaining decisions in terms of interpretable patterns, such as unusual transaction sequences, deviations from historical behavior, or known fraud typologies, rather than abstract feature weights or scores. For example, in current financial crime detection, XAI typically surfaces technical signals such as “Transaction\_Amount” and “IP\_Distance\_KM” [34]. A human-centric approach will provide explanations that would identify a particular transaction’s behavior as “Suspicious Location” or “Money Mule” fraud typology. This shift aligns the system’s output with the analyst’s existing mental models.

LLMs may offer opportunities to generate concept-level explanations due to their strong natural language generation capabilities and ability to incorporate contextual information. Their extensive pre-training on large-scale datasets also enables a degree of generalization without requiring additional domain-specific input from end users [49]. However, the use of LLMs in this context must be carefully calibrated. Affected users may already be frustrated, anxious, or time-constrained when transactions are blocked, and overly verbose or poorly scoped explanations risk increasing cognitive burden rather than supporting understanding [10]. Moreover, privacy and confidentiality constraints are

particularly salient in banking and financial services, further limiting how explanations can be generated and delivered. These considerations highlight the need for explanation designs that balance expressiveness with restraint and that are sensitive to users’ cognitive load and emotional state in high-stakes financial contexts.

### 3.3 Harnessing Agentic AI in Financial Crime Detection

Regulators increasingly expect financial institutions and technology companies to maintain financial crime detection systems that respond rapidly to emerging fraud and money laundering typologies, ideally approaching real-time detection [11]. In practice, this expectation creates significant operational challenges for those responsible for detection. When new typologies emerge, analysts must interpret evolving criminal behaviors, decompose them into logical conditions, and translate these insights into new or modified detection rules. This process is highly manual, resource-intensive, and dependent on scarce expertise spanning both illicit activity and normal customer behavior. As a result, detection systems are often updated reactively, leaving defenders structurally behind attackers who can adapt more quickly and opportunistically.

To narrow this gap, fraud detection vendors are integrating autonomous agents into detection workflows to facilitate continuous, systematic adaptation. Unlike traditional models, agentic systems can ingest transaction patterns, alert outcomes, and investigation results over time to identify gaps in existing rule coverage. When novel patterns emerge, these agents can synthesise targeted rule modifications, aligning detection speed with regulatory expectations while reducing manual effort [22]. We have begun to observe this shift in industry deployments. For instance, Mastercard’s Agent Pay initiative illustrates how credit card networks are beginning to embed agentic AI into their fraud management ecosystems. Agentic workflows are characterized by (1) autonomy, (2) goal complexity, (3) environmental adaptability and (4) decision-making, often within multi-agent architectures [2], introduce additional layers of complexity and raise a question for explainable AI (XAI): does the rise of agentic systems fundamentally alter XAI requirements, or does it instead amplify the importance of existing principles such as interactivity and transparency? We argue that while these systems increase the difficulty of XAI implementation, requiring analysts to trace and decompose decisions across multiple agents and steps, they do not fundamentally change its underlying principles. Rather, existing XAI capabilities become more critical in financial crime detection, as they enhance analysts’ understanding of agent behavior and support the early identification of malfunction or compromise by enabling reporting of anomalous behavioral shifts. Although empirical studies on the effects of agentic AI on explainability and decision effectiveness in financial crime detection remain limited, insights can be drawn from related domains. For example, agent-based approaches developed for detecting social engineering attacks [26, 37], where threat patterns evolve rapidly and systems must be continuously updated, may offer useful strategies for adapting XAI to dynamic financial crime contexts.

## 4 Conclusion

XAI is critical for supporting trust, fairness, accountability, and transparency in high-stakes domains such as financial crime detection. However, much of existing XAI research remains insufficiently grounded in the operational realities and stakeholder needs of this domain, often prioritizing technical properties of explanations over their practical use and impact. We identify new research directions in XAI for financial crime detection that center on understanding stakeholder needs, designing human-centric explanations, while keeping the risks and limitations of XAI methods in mind. XAI should be viewed as one component within a broader ecosystem of tools that enable human questioning, clarification, and oversight. Rather than striving for a single “optimal” explanation, XAI should be only one part of a broader portfolio of stakeholder-centered approaches that includes platforms for users and affected parties to

challenge AI decisions, enable recourse, and provide feedback [6]. Addressing these challenges requires interdisciplinary collaboration across human-computer interaction, psychology, and law, as well as with practitioners and regulators. Building on prior work presented at HCXAI and emerging XAI research in the financial sector, this position paper aims to serve as a first step toward developing AI systems that are not only explainable but also adapted to human and organizational contexts.

## References

- [1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanahalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018-04-21) (*CHI '18*). Association for Computing Machinery, 1–18. doi:10.1145/3173574.3174156
- [2] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B. Divya. 2025. Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access* 13 (2025), 18912–18936. doi:10.1109/ACCESS.2025.3532853
- [3] Monica Violeta Achim and Sorin Nicolae Borlea. 2020. *Economic and Financial Crime : Corruption, Shadow Economy, and Money Laundering*. Springer. <https://link.springer.com/book/10.1007/978-3-030-51780-9>
- [4] Eman Alashwali, Ragashree Mysuru Chandrashekar, Mandy Lanyon, and Lorrie Faith Cranor. 2024. Detection and Impact of Debit/Credit Card Fraud: Victims' Experiences. In *Proceedings of the 2024 European Symposium on Usable Security* (New York, NY, USA, 2024-11-20) (*EuroUSEC '24*). Association for Computing Machinery, 235–260. doi:10.1145/3688459.3688464
- [5] Sule Anjomshoe, Kary Främling, and Amro Najjar. 2020. *Explanations of Black-Box Model Predictions by Contextual Importance and Utility*. arXiv.org. doi:10.1007/978-3-030-30391-4\_6
- [6] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*. arXiv:1909.03012 [cs] doi:10.48550/arXiv.1909.03012
- [7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bénéttot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (2020), 82–115. doi:10.1016/j.inffus.2019.12.012
- [8] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018-04-21) (*CHI '18*). Association for Computing Machinery, 1–14. doi:10.1145/3173574.3173951
- [9] Ruth M. J. Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19* (2019-07). International Joint Conferences on Artificial Intelligence Organization, 6276–6282. doi:10.24963/ijcai.2019/876
- [10] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Navid Nobani, and Andrea Seveso. 2024. *XAI Meets LLMs: A Survey of the Relation between Explainable AI and Large Language Models*. arXiv:2407.15248 [cs] doi:10.48550/arXiv.2407.15248
- [11] Saaniya Chugh and Aditya Vilas Deshpande. 2025. *Opportunities and Challenges of Agentic AI in Finance*. Social Science Research Network:5538799 doi:10.2139/ssrn.5538799
- [12] Federal Trade Commission. 2025. *New FTC Data Show a Big Jump in Reported Losses to Fraud to \$12.5 Billion in 2024*. Federal Trade Commission. <https://www.ftc.gov/news-events/news/press-releases/2025/03/new-ftc-data-show-big-jump-reported-losses-fraud-125-billion-2024>
- [13] Dan Conway, Ronnie Taib, Mitch Harris, Shlomo Berkovsky, Kun Yu, and Fang Chen. 2017. A Qualitative Investigation of Bank Employee Experiences of Information Security and Phishing. In *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security* (USA, 2017-07-12) (*SOU'PS '17*). USENIX Association, 115–129.
- [14] Irina Dimitrova, Peter Öhman, and Darush Yazdanfar. 2021. Barriers to Bank Customers' Intention to Fully Adopt Digital Payment Methods. *International Journal of Quality and Service Sciences* 14, 5 (2021), 16–36. doi:10.1108/IJQSS-03-2021-0045
- [15] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2019-03-17) (*IUI '19*). Association for Computing Machinery, 275–285. doi:10.1145/3301275.3302310
- [16] Finale Doshi-Velez and Been Kim. 2017. *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv:1702.08608 [cs] doi:10.48550/arXiv.1702.08608
- [17] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems. (2021). doi:10.48550/ARXIV.2101.04719
- [18] Mica R. Endsley. 2017. *From Here to Autonomy*. Human factors. doi:10.1177/0018720816681350
- [19] European Central Bank. 2025. Joint EBA-ECB Report on Payment Fraud: Strong Authentication Remains Effective but Fraudsters Are Adapting. (2025). <https://www.ecb.europa.eu/press/pr/date/2025/html/ecb.pr251215-e133d9d683.en.html>
- [20] Nishani Fernando, Bahareh Nakisa, Adnan Ahmad, and Mohammad Naim Rastgoo. 2025. *Adaptive XAI in High Stakes Environments: Modeling Swift Trust with Multimodal Feedback in Human AI Teams*. arXiv:2507.21158 [cs] doi:10.48550/arXiv.2507.21158
- [21] International Monetary Fund. 2001. Financial System Abuse, Financial Crime and Money Laundering—Background Paper. <https://www.imf.org/external/np/ml/2001/eng/021201.htm>
- [22] Nitish Jaipuria, Lorenzo Gatto, Zijun Kan, Shankey Poddar, Bill Cheung, Diksha Bansal, Ramanan Balakrishnan, Aviral Suri, and Jose Estevez. 2025. *CASE: An Agentic AI Framework for Enhancing Scam Intelligence in Digital Payments*. arXiv:2508.19932 [cs] doi:10.48550/arXiv.2508.19932
- [23] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 2021. *If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques*. arXiv:2103.01035 [cs] doi:10.48550/arXiv.2103.01035

- [24] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2016-05-07) (*CHI '16*). Association for Computing Machinery, 2390–2395. doi:10.1145/2858036.2858402
- [25] Maya Krishnan. 2020. Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology* 33, 3 (2020), 487–502. doi:10.1007/s13347-019-00372-9
- [26] Nir Kshetri. 2025. Transforming Cybersecurity with Agentic AI to Combat Emerging Cyber Threats. *Telecommunications Policy* 49, 6 (2025), 102976. doi:10.1016/j.telpol.2025.102976
- [27] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence* 296 (2021), 103473. doi:10.1016/j.artint.2021.103473
- [28] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. (2020). doi:10.48550/ARXIV.2001.02478
- [29] Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why Does My Model Fail? Contrastive Local Explanations for Retail Forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2020-01-27) (*FAT'20*). Association for Computing Machinery, 90–98. doi:10.1145/3351095.3372824
- [30] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2017-12-04) (*NIPS'17*). 4768–4777. <https://dl.acm.org/doi/10.5555/3295222.3295230>
- [31] Ramesh Manuvinakurike, Emanuel Moss, Elizabeth Anne Watkins, Saurav Sahay, Giuseppe Raffa, and Lama Nachman. 2025. *Thoughts without Thinking: Reconsidering the Explanatory Value of Chain-of-Thought Reasoning in LLMs through Agentic Pipelines*. arXiv:2505.00875 [cs] doi:10.48550/arXiv.2505.00875
- [32] Andreas Matthias. 2004. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology* 6, 3 (2004), 175–183. doi:10.1007/s10676-004-3422-1
- [33] Carlo Metta, Riccardo Guidotti, Yuan Yin, Patrick Gallinari, and Salvatore Rinzivillo. 2022. Exemplars and Counterexemplars Explanations for Skin Lesion Classifiers. In *Front. Artif. Intell. Appl.* (2022), Vol. 354. IOS Press BV, 258–260. doi:10.3233/FAIA220209
- [34] Eleanor Ruth Mill, Wolfgang Garn, Nicholas F. Ryman-Tubb, and Christopher Turner. 2023. Opportunities in Real Time Fraud Detection: An Explainable Artificial Intelligence (XAI) Research Agenda. *International Journal of Advanced Computer Science and Applications* 14, 5 (2023), 1172–1186. doi:10.14569/IJACSA.2023.01405121
- [35] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38. doi:10.1016/j.artint.2018.07.007
- [36] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2018. *Explaining Explanations in AI*. arXiv.org. doi:10.1145/3287560.3287574
- [37] Anwar Mohammed. 2025. Agentic AI as a Proactive Cybercrime Sentinel: Detecting and Deterring Social Engineering Attacks. *Journal of Data and Digital Innovation (JDDI)* 2, 2 (2025), 109–117. <http://datalensjournal.com/index.php/JDDI/article/view/18>
- [38] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User-Adapted Interaction* 27, 3 (2017), 393–444. doi:10.1007/s11257-017-9195-0
- [39] Wolter Pieters. 2011. Explanation and Trust: What to Tell the User in Security and AI? *Ethics and Information Technology* 13, 1 (2011), 53–64. doi:10.1007/s10676-010-9253-3
- [40] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. *Stakeholders in Explainable AI*. arXiv:1810.00184 [cs] doi:10.48550/arXiv.1810.00184
- [41] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2020-01-27) (*FAT'20*). Association for Computing Machinery, 33–44. doi:10.1145/3351095.3372873
- [42] Nidhi Rastogi, Shirid Pant, Devang Dhanuka, Amulya Saxena, and Pranjal Mairal. 2025. *Too Much to Trust? Measuring the Security and Cognitive Impacts of Explainability in AI-Driven SOCs*. arXiv:2503.02065 [cs] doi:10.48550/arXiv.2503.02065
- [43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. doi:10.48550/ARXIV.1602.04938
- [44] Zahra Sadeghi, Roohallah Alizadehsani, Mehmet Akif Cifci, Samina Kausar, Rizwan Rehman, Priyakshi Mahanta, Pranjal Kumar Bora, Ammar Almasri, Rami S. Alkhawaldeh, Sadiq Hussain, Bilal Alatas, Afshin Shoeibi, Hossein Moosaei, Milan Hladik, Saeid Nahavandi, and Panos M. Pardalos. 2024. A Review of Explainable Artificial Intelligence in Healthcare. *Computers and Electrical Engineering* 118 (2024), 109370. doi:10.1016/j.compeleceng.2024.109370
- [45] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. arXiv:2305.04388 [cs] doi:10.48550/arXiv.2305.04388
- [46] Roy Wedge, James Max Kanter, Santiago Moral Rubio, Sergio Iglesias Perez, and Kalyan Veeramachaneni. 2017. *Solving the "False Positives" Problem in Fraud Prediction*. arXiv:1710.07709 [cs] doi:10.48550/arXiv.1710.07709
- [47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv:2201.11903 [cs] doi:10.48550/arXiv.2201.11903



- [48] Jarrod West and Maumita Bhattacharya. 2016. Intelligent Financial Fraud Detection: A Comprehensive Review. *Computers & Security* 57 (2016), 47–66. doi:10.1016/j.cose.2015.09.005
- [49] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. *Explainability for Large Language Models: A Survey*. arXiv:2309.01029 [cs] doi:10.48550/arXiv.2309.01029