Trust in Transparency: How Explainable AI Shapes User Perceptions

Abstract

This study explores the integration of contextual explanations into AI-powered loan decision systems to enhance trust and usability. While traditional AI systems rely heavily on algorithmic transparency and technical accuracy, they often fail to account for broader social and economic contexts. Through a qualitative study, I investigated user interactions with AI explanations and identified key gaps, including the inability of current systems to provide context. My findings underscore the limitations of purely technical transparency and the critical need for contextual explanations that bridge the gap between algorithmic outputs and real-world decision-making. By aligning explanations with user needs and broader societal factors, the system aims to foster trust, improve decision-making, and advance the design of human-centered AI systems.

Keywords: Explainable AI (XAI), Human-AI Interaction (HAI), Sociotechnical Systems, Algorithmic Transparency, Contextual Explanations.

1 Introduction

We live in an age of technological acceleration. The telephone took 78 years to reach 50 million users, while radio took 38 years.[37] In stark contrast, ChatGPT achieved the same milestone in just two months. [41]. This rapid adoption underscores a seismic shift in how society integrates new technologies. Historically, the ability to absorb economic and social changes caused by technological breakthroughs has taken time, allowing for trust to build naturally through iterative use and understanding [32] However, the accelerated pace at which AI systems are being deployed has disrupted this historical pattern. Today, we see AI systems being incorporated into critical decision-making roles such as finance[28], college admissions[34], criminal justice[42] and the banking industry.[13]

To safety adopt technology, you need to trust it.[11] A critical component of this trust-building process is explainability. [9][36] For previous technological breakthroughs, this trust-explainability relationship was fostered through transparency and accountability. Engineers understood the systems they built and could explain them to stakeholders.[26] Today, with the rise of opaque AI models, this foundational understanding is often missing. Even the engineers behind these systems frequently lack insight into how decisions are made, leaving end-users and stakeholders blind to critical processes.[21][2]

Responding to this issue, the field of Explainable AI (XAI) has progressed at a rapid clip. [30] It has given us algorithmic approaches to generate explanations of how an AI system behaves and makes decisions.[12] The field has, however, been criticized for having a myopic focus.[20] [38] As most XAI techniques are designed by XAI researchers

for other XAI researchers, it excludes those who are actually affected by the AI system. [4] Research has shown that the explanations preferred by XAI researchers do not translate well to other individuals who are actually affected by these systems.[39] [7] As with most poorly deployed systems, it tends to affect the most marginalized within society.[38] Explanations as a construct are usually more effective when looked at through a socio-technical lens. Explanation is first and foremost a shared meaning-making process that occurs between an explainer and an explainee. This process is dynamic to the goals and changing beliefs of both parties.[16]

In response to these criticisms, there have been several studies, including [35], [44], [23]that have attempted to extend user centric AI development in order to develop paradigms that can help define certain benchmarks in creating AI systems that are able to build trust. This has helped create a new field, Human-AI Interaction (HAI). This new subfield has tackled the problem of building trust by bridging work between AI, HCI and work in critical theory such as [5] that have given insights into this issue. Attempts have been made in order to conclusively design frameworks that address the explainability-trust framework. [22] attempted to use critical theory to analyze and critique the social and ethical implications of explanations generated by AI systems. This approach emphasizes the role of context and power dynamics in determining what constitutes a "reasonable" explanation. On the other hand, [27] provided a foundational understanding of explanations by leveraging insights from cognitive and social sciences, focusing on how humans generate and interpret explanations in contrastive and causal terms. Building on these perspectives, [29] defined the COP-12 metrics, which offer a structured framework for evaluating the quality of AI explanations. These have broad themes of Content, Presentation and User Focus. These metrics emphasize key dimensions such as clarity, relevance, fidelity, and utility, ensuring that explanations are not only accurate but also meaningful and actionable for end-users.

In spite of these efforts, most of the research in the HAI space does not touch users that have no context of the system they are going to be affected by. Most current studies such as [3][16] use experts in their respective fields. Surveys of papers such as [29][35] are still biased towards expert opinions. In this paper, we look at the current explanation paradigms in the HAI field and attempt to see if they still hold weight when evaluated by end-users who have very limited understanding of AI systems as well as minimal contextual knowledge. Or more broadly, how useful is the current state-of-the-art in explanations within the HAI field in this new context? In summary, our contributions are:

- RQ1: How do different types of AI explanations (e.g., example-based, rule-based) shape users' initial impressions and perceived trust during their interaction with the AI system?
- RQ2: What elements of these explanations are most influential in building or undermining users' trust during their interactions with a AI system?
- Reaffirming whether the current metrics used to evaluate HAI are effective in diverse real-world contexts and what perspectives are missing

2 Related Work

We will start with a review of the XAI field and how a new sub field called Human AI Interaction (HAI) has arisen. We'll also review the benefits of looking at AI development through a sociotechnical scale.

2.1 Explainable AI (XAI)

Explainable AI (XAI) aims to uncover how AI models behave and to communicate these behaviors to users in a way that fosters understanding.[2] The nature of XAI work varies depending on the specific objectives, whether it's diagnosing model behavior or providing actionable insights to end-users.[1] Researchers in this field often develop diagnostic explanations, designed to help practitioners and developers identify and address model issues. In contrast, actionable explanations are more suited to non-technical users, enabling them to make informed decisions without requiring a deep understanding of the model's inner workings.[21]

To address the needs of broader audiences, XAI research has introduced post-hoc explanations, such as counterfactual explanations [24], which aim to elucidate model behavior without relying on diagnostic tools. These explanations attempt to bridge the gap for users who lack the technical capability to interpret complex model outputs. However, recent findings have shown that explainability is audience-dependent rather than model-deterministic; what resonates with one user may fail to convey meaningful insights to another. This highlights the critical need for tailoring explanations to specific user groups.[10]

Despite these advancements, significant challenges persist in XAI research. Many common XAI techniques have been tested only in controlled laboratory settings with user research [14], limiting their generalizability. In practice, these techniques often struggle to effectively convey explainability to diverse user populations. Moreover, the relationship between explainability and user trust has shown mixed results, with XAI techniques frequently falling short in fostering trust among users.[11] [15] Achieving both explainability and trustworthiness in real-world systems remains a persistent challenge.

Another issue lies in the sociotechnical aspect of XAI. Much of the current research focuses on the deployment of AI systems in closed business environments, analyzing their behavior within these constrained settings. [3]There is limited exploration of how XAI systems perform when deployed at scale in large, open environments, where diverse user needs and societal implications come into play.

2.2 Human-AI Interaction (HAI)

Current HAI techniques focus on improving user interaction and fostering trust through mechanisms such as explainability, interpretability, and personalized design.[33] Methods like Explainable AI (XAI) aim to make AI systems more transparent by providing insights into how decisions are made, while techniques such as interactive visualizations and natural language explanations enhance user engagement and understanding. Personalization is increasingly emphasized, with AI systems adapting to individual user needs, contexts, and expertise levels. [17] However, these approaches face significant critiques. Many explanations remain too complex, failing to account for diverse user capabilities

and cognitive loads.[25] The "black-box" problem persists, with opaque algorithms making it difficult for even developers to fully understand AI decision-making processes. Furthermore, current HAI techniques often lack empirical grounding in psychological or sociological research, resulting in explanations that may not align with how users naturally process information.[27] Ethical concerns, such as potential bias in explanations and the manipulation of user perceptions, also highlight the need for critical evaluation and more robust, inclusive design frameworks. As AI systems increasingly take on high-stakes roles, addressing these critiques is essential for building trustworthy, human-centered solutions.[40] Although HAI is growing in a more inclusive direction, the vast amount of it's research is still focused around the computer science or domain specific spaces. Wide ranging surveys such as [29] [43][35] [31] all showcase this concentration. In particular [29] from which the COP-12 metrics were derived shows that of the nearly 400 papers under review, 22% of them only contain user studies. Out of that , only 23% of the remaining are focused on non-domain experts. This skews reported metrics which fails to take into account perspectives that are completely outside of the space.

2.2.1 COP-12

The COP-12 framework, introduced by Nauta et al[29]., is a structured set of metrics designed to evaluate the quality of AI explanations. It identifies 12 key dimensions grouped into three main categories: Content, Presentation, and User-Focused Metrics. Content metrics (e.g., correctness, completeness, consistency) assess the factual accuracy and comprehensiveness of explanations. Presentation metrics (e.g., compactness, composition) evaluate how explanations are structured and delivered to users. Finally, user-focused metrics (e.g., context, coherence, controllability) emphasize the alignment of explanations with user needs and their ability to interact with and understand the system. By providing a comprehensive evaluation framework, COP-12 aims to bridge technical accuracy with user-centric design principles, making it particularly suited for assessing the explainability and trustworthiness of AI systems.

2.3 Sociotechnical Perspectives

Sociotechnical approaches to AI emphasize embedding systems within their broader social and organizational contexts, arguing that technological solutions alone are insufficient for addressing complex real-world challenges. Such approaches integrate human and organizational factors, fostering systems that are both technically effective and socially meaningful.

Frameworks like *Critical Technical Practice (CTP)* [5] and *Value-Sensitive Design (VSD)* [45] exemplify this perspective. CTP, as proposed by Agre, calls for a reflective critique of the epistemic and methodological assumptions underlying AI development. It encourages researchers to question dominant algorithmic paradigms and consider alternative designs that prioritize human values and contextual relevance. Similarly, VSD integrates stakeholder perspectives early in the design process to ensure that the systems align with the ethical and practical needs of diverse user groups.

These perspectives challenge the dominance of algorithmic formalism [19] the tendency to abstract AI solutions away from their real-world context. For instance, studies have highlighted how neglecting sociotechnical factors can lead to algorithmic interventions that perpetuate biases, exacerbate inequalities, or fail to address user needs effectively. By

extending abstraction boundaries to include social, cultural, and organizational factors, sociotechnical approaches aim to mitigate these issues, ensuring that AI systems are not only transparent but also trustworthy and fair.

The emphasis on sociotechnical integration aligns with broader calls for *localized, context-sensitive solutions* in AI development. Rather than relying on scalable, one-size-fits-all models, these approaches advocate for systems tailored to specific social and organizational settings. This localization ensures that the systems resonate with the lived experiences of their users, addressing unique challenges and opportunities within their deployment contexts.

Ultimately, sociotechnical perspectives argue for a paradigm shift in AI design—moving from algorithm-centric approaches to human-centered systems that are reflexive, inclusive, and deeply embedded within the social fabrics they aim to serve.

3 Methods

3.1 Recruitment

This study was conducted within my neighborhood and extended friend circle, leveraging snowball sampling to recruit participants.[18] Snowball sampling was chosen to efficiently identify individuals who fit the study's criteria, given the preliminary nature of the research and the constraints of time. Due to these constraints, flyer recruitment was not employed, as the goal was to quickly gather participants for an exploratory investigation.

Participants were specifically chosen based on their limited prior knowledge of AI systems, ensuring that the study focused on individuals with minimal exposure to or understanding of such technologies. Recruitment was verified through unscripted discussions, during which the following factors were assessed:

- **Participant Identification:** Verification of eligibility for the study.
- **Comfort Level with AI Systems:** Ensuring participants had no intimate familiarity with AI systems.
- **AI System Familiarity:** Determining the type of AI systems participants were most comfortable or familiar with, if any.

Participants were asked to consent to a 30-minute discussion, during which their interactions with the system were observed and analyzed. Participants provided informed consent before participating in the study and their data has been annonymized. A total of seven participants were recruited. Two of them had some understanding of AI systems. Three of the remaining had used ChatGPT in the last six months and hence approached the AI system with that knowledge in hand. Table 1 contains the participant ID, job, and AI knowledge.

Participant ID	Job	AI Knowledge
P1	Data Science Student	Medium
P2	Electrical Engineering Student	Low
P3	Data Science Student	Medium
P4	Business School Major	Low
P5	Retired	Low
P6	Graphic Designer	Low
P7	Marketing	Low

Table 1: Participant Codes

3.2 Technical Design

The system for this study was developed to evaluate the effects of various AI explanation types on user trust and understanding. The design integrates cutting-edge technologies to provide a flexible, interactive environment for qualitative analysis. The system consists of a React front end, an LLM-powered backend, and a Python-based explanation solver, with the following components:

Frontend Design The frontend of the system was implemented using React, enabling a user-friendly and interactive interface. This interface presents AI model predictions alongside different types of explanations. It allows users to explore explanations through dynamic explanations and textual descriptions. it also Captures user feedback and interaction data for subsequent analysis.

Backend and Explanation Engine The backend architecture leverages an LLM (Llama 3.2B)[6] to enhance the system's ability to provide contextually relevant and dynamic explanations. The backend is designed to generate contextual information. The Llama 3.2B model generates natural language summaries of model predictions and contextualizes explanations based on user profiles and tasks. The model adapts explanations dynamically to fit the specific scenario or user input, tailoring outputs to enhance user comprehension.

Explanation Solver The system employs a Python-based solver to compute and present interpretable explanations. A key component is the SHAP (SHapley Additive exPlanations) package, which is used to calculate feature importance scores for each model prediction. it can then visualize how individual features contribute to the output, providing diagnostic insights for technical users and finally generate concise, non-technical explanations for users without deep technical expertise.

3.3 AI Interaction

Participants interacted with an AI-powered loan decision system by simulating loan applications. They provided inputs such as age, sex, requested credit amount, and income. Based on these inputs, the system would either approve or reject the loan request. After each decision, the system could provide an explanation detailing the reasoning behind its decision, while other times, no explanation was given. Participants were encouraged to submit multiple loan requests with varying inputs to explore the system's decision-making

logic. Participants interacted with the system ten times with a different explanation type each time. Participants were encouraged to reflect on why the system was behaving in this manner and if the explanations generated were reasonable.

3.3.1 Explanation Types

The system generates and displays four types of explanations:

- **No Explanation:** No explanation is generated on the screen.
- **Basic Explanation:** Feature importance explanations.
- **Detailed Explanation:** Contextual information.
- **Interactive Explanation:** Dynamic user queries.

3.4 The Interview Process

The semi structured interviews were conducted both online over zoom and in-person. The interview had three main parts:

1. **Part 1:** A casual discussion about the participants' prior experiences with AI systems, including their perceptions and challenges, serving as a foundation for understanding their familiarity and expectations.
2. **Part 2:** Structured questions directly aligned with the Cop-12 metrics, targeting key aspects such as correctness, context, and confidence in explanations. While this section followed a relaxed framework to ensure all relevant metrics were addressed, the conversation was deliberately guided to encourage participants to reflect deeply on these aspects.
3. **Part 3:** A relaxed, post-interview style, allowing participants to share their broader thoughts and fill out related feedback forms, creating an informal atmosphere to elicit candid responses.

   The entire Interview protocol can be viewed in the appendix.

   the results were coded so that they aligned as much as possible with the already existing Cop-12 metrics and attempted to see if there were themes that are not captured previously.

3.5 Qualitative Analysis

The qualitative analysis followed a structured approach to ensure the systematic identification of themes aligned with the Cop-12 metrics. The process began with familiarization, where interview transcripts were reviewed to understand the broad patterns and themes emerging across participant responses. Next, initial coding involved labeling specific transcript segments with descriptive codes, guided by the Cop-12 metrics. For example, one participant's statement, "The explanation didn't reflect what the system actually did" (P1), was coded as "incorrect explanation" under the Cop-12 theme of correctness. Following this, thematic coding merged similar codes into sub-themes that remained aligned with the Cop-12 framework. For

instance, codes such as "Inconsistent Outcomes" and "Incorrect Explanation" were consolidated under the sub-theme "Content Issues," corresponding to the Cop-12 theme of content.The process of theme development involved refining and defining these sub-themes as they began to emerge more distinctly. Finally, during validation, the themes were further refined to ensure coherence and alignment with both the data and the Cop-12 metrics. This iterative process ensured that the analysis was grounded in participant responses while maintaining relevance to the established theoretical framework.[8] Once the codes were matched with the already existing codes, further analysis could be done on the remaining themes.

4 Findings

The findings from the study reveal that most participant responses align well with the COP-12 framework, particularly in terms of its emphasis on content, presentation, and user-focused metrics. However, certain areas of participant feedback expose gaps in COP-12, suggesting that the framework could benefit from expanded dimensions to fully capture user needs and experiences.

4.1 Research Question 1

How do different types of AI explanations shape users' initial impressions and perceived trust during their interaction with the AI system?

Participants' impressions of the AI system were strongly influenced by the type of explanation provided and its alignment with their expectations. Explanations that demonstrated consistency and correctness were most likely to build trust. For example, P1 noted, *"When I gave 30k as income it gave me acceptance, but when I dropped it to 25k it rejected me,"* highlighting the importance of consistency in maintaining user trust. Similarly, P3 expressed skepticism when explanations lacked sufficient detail to clarify key factors: *"I guess I understood that the income amount has greater weight compared to sex, but I'm not sure how much the weight is."* These findings align with the correctness and covariate complexity sub-themes within the Content dimension of COP-12.

The type of explanation also played a role in shaping user impressions based on their level of expertise. Novice users preferred simple, contextual explanations, while participants with more experience appreciated technical details. Interactive explanations, which allowed users to query and clarify the AI's decisions, stood out as particularly effective. P7's question, *"Can we ask it questions?"* reflects the growing expectation for interactive systems, tying directly to the controllability sub-theme under COP-12's User-Focused Metrics. However, inconsistencies in how explanations adapted to user queries and inputs highlighted potential gaps in COP-12's framework for evaluating dynamic, interactive systems.

4.2 Research Question 2

What elements of these explanations are most influential in building or undermining users' trust during their interactions with this AI system?

Participants identified several elements of AI explanations that influenced their trust, with transparency and contextual relevance emerging as critical factors. Transparency, particularly regarding confidence indicators, played a significant role. P4 questioned, *"Is an accuracy score of 85% even good?"* and further asked, *"Is the confidence level for the final result or the explanation?"* Similarly, P7 added, *"If the confidence level is high and the explanation is bad, is the system good or bad? What does the confidence level actually mean here?"* These quotes highlight the confidence sub-theme under COP-12, where users expect clear and actionable explanations of confidence values. However, the lack of clarity around confidence metrics suggests that COP-12 could benefit from more detailed guidance on presenting this information.

Contextual relevance was also crucial in fostering trust. Users valued explanations tailored to their specific decisions over generic outputs. P5 observed, *"Real life is more than numbers, I would just talk to someone at the Bank"* emphasizing the need for explanations to account for real-world factors and specific decision contexts. While COP-12 captures context as a sub-theme, participants' feedback points to gaps in the framework's ability to evaluate how explanations adapt dynamically to user needs or queries.

A couple participants highlighted issues with accessibility and usability that undermined their trust. For example, P3 commented, *"There is no way to control the length of the response the way I can do it in ChatGPT,"* reflecting the importance of compactness in ensuring explanations are concise and user-friendly. This feedback underscores the need for AI explanations to balance detail with simplicity. Accessibility challenges also extended to users' ability to interpret technical metrics, such as accuracy. P1 remarked, *"I don't think 85% is very high accuracy,"* while P3 added, *"I think you should get the accuracy a little higher before deploying the system."* These statements suggest that while correctness and confidence are covered in COP-12, the framework does not explicitly account for how such metrics should be presented to ensure accessibility and usability for diverse user groups.

4.3 Tie-in with COP-12 Metrics and Limitations

The findings demonstrate that the COP-12 framework effectively captures many dimensions critical to explainability, including correctness, consistency, compactness, and context. For example:

- **Correctness and Consistency:** P1's remark about the system's inconsistent handling of similar income inputs highlights the importance of these subthemes in fostering trust.
- **Compactness:** P3's frustration with verbose explanations underscores the need for concise presentation formats.
- **Confidence:** P4 and P7's concerns about unclear confidence metrics point to the need for transparency in this area.

However, the findings also reveal areas where COP-12 falls short:

- **Fairness:** Participants questioned the ethical implications of system decisions, such as P4's remark about the relevance of accuracy scores. COP-12 does not explicitly evaluate fairness or user perceptions of equity in AI outcomes.
- **Accessibility:** P3's comments on explanation length and usability, along with P1's and P3's remarks on interpreting accuracy scores, highlight a gap in COP-12's ability to evaluate whether explanations are accessible and actionable for users with varying levels of expertise.
- **Adaptability:** While COP-12 addresses controllability, it does not fully account for the dynamic and interactive nature of explanations that participants, like P7, found valuable.

5 Discussion

In the discussion section I will describe two broad themes that have come out of our analysis.

Algorithmic Transparency Is Not Enough

The findings from our study emphasize that while algorithmic transparency is a necessary step toward fostering trust in AI systems, it is far from sufficient. Participants frequently expressed skepticism about the relevance and clarity of purely technical explanations, even when these were accurate and detailed. For instance, Participant P4's comment, *"Is an accuracy score of 85% even good?"* highlights the limitations of presenting technical metrics without sufficient contextual framing. Additionally, the lack of clarity in confidence indicators, as illustrated by P7's confusion over what confidence scores represent, suggests that transparency efforts often fail to translate into actionable understanding for users.

Interactive explanations showed promise in bridging some of these gaps, particularly for users who were able to engage with the system through queries. However, even these explanations struggled to adapt dynamically to user needs or clarify inconsistencies in the AI's behavior. For example, P1 noted the inconsistent handling of similar income inputs, which undermined their trust in the system. This highlights a broader issue: algorithmic transparency, as defined by metrics like correctness and compactness under COP-12, does not inherently ensure that users find the system reliable or fair.

The findings underscore the need for transparency approaches that go beyond algorithmic details. These must address user-centric dimensions such as accessibility, interpretability, and adaptability. Without these considerations, transparency efforts risk alienating users or, worse, fostering a false sense of trust in systems that fail to account for real-world complexities.

The Need for a Wider Social Context

A recurring theme in participant feedback was the disconnect between the AI explanations provided and the broader social, economic, and individual contexts in which users make decisions. Participants like P5 underscored the limitations of numerical and technical outputs, remarking, *"Real life is more than numbers, I would just talk to someone at the Bank."* This sentiment reflects a critical gap in the design of current explanation systems: their inability to contextualize decisions within the lived realities of users.

Contextual explanations, which incorporate factors such as local economic conditions and individual circumstances, were highlighted as pivotal for fostering trust. However, the findings also reveal that existing frameworks like COP-12 inadequately address this need. While the framework includes *context* as a dimension, it does not fully evaluate how explanations adapt dynamically to user-specific scenarios or societal factors. This limitation became evident in the feedback from participants who sought explanations that were not just technically accurate but also socially meaningful.

Moreover, participants' concerns about fairness and accessibility point to the need for explanation systems that engage with ethical and sociotechnical dimensions of AI. P4's questioning of the relevance of accuracy scores, for example, underscores the importance of designing systems that account for perceptions of equity and inclusivity. Similarly, P3's frustration with verbose and inaccessible explanations highlights the need for systems that prioritize user-centric design over purely technical objectives.

By embedding explanations within a wider social context, AI systems can move beyond the limitations of algorithmic formalism to become truly human-centered. This involves not only tailoring explanations to individual users but also addressing systemic factors that influence trust, such as power dynamics, cultural norms, and economic inequalities. Only by adopting such a sociotechnical perspective can AI systems meaningfully align with the needs and expectations of diverse user groups, fostering trust in high-stakes environments.

6 Limitations and Future Work

There were several limitations on the study. The participant size was very small and there was no saturation in the codes. Findings from this study should not be generalized. From a technical stand point the way the interaction worked was substandard. A lot of problems might have been alleviated in case of better hardware or more investment into the LLM.

Future work will include running the experiments again with a larger sample size and a more elaborate technical set up.

7 Conclusion

This study underscores the limitations of current approaches to explainability in AI systems, particularly those focused narrowly on algorithmic transparency. While technical transparency is essential, it does not inherently foster trust or ensure usability, as evidenced by participant feedback highlighting issues with confidence indicators, consistency, and accessibility. The findings reveal that effective explanations must extend beyond algorithmic details to address user-centric and sociotechnical dimensions.

Contextual explanations that incorporate societal, economic, and individual factors emerged as critical for fostering trust. By aligning explanations with the lived realities of users and addressing systemic factors such as fairness and accessibility, AI systems can move beyond their current limitations. This study advocates for a paradigm shift toward sociotechnical perspectives in AI design, emphasizing the importance of creating human-centered systems that are not only transparent but also equitable and meaningful in real-world contexts. Future work should aim to refine frameworks like COP-12 to account for dynamic, interactive, and contextually rich explanations, ensuring that AI systems effectively meet the diverse needs of their users.

References

[1] Arrieta, a. b., díaz-rodríguez, n., del ser, j., bennetot, a., tabik, s., barbado, a., ... herrera, f. (2020). explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. information fusion, 58, 82-115.

[2] Mccarthy, s. (2022). development of an explainability scale to evaluate explainable artificial intelligence (xai) methods.

[3] Q. vera liao, d. gruen, and s. miller, "questioning the ai: informing design practices for explainable ai user experiences," in chi'20: Proceedings of the 2020 chi conference on human factors in computing systems, pp. 1–15, new york, ny, usa, 2020.

[4] J. Aechtner, L. Cabrera, D. Katwal, P. Onghena, D. P. Valenzuela, and A. Wilbik. Comparing user perception of explanations developed with xai methods. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, Padua, Italy, 2022. IEEE.

[5] Philip E. Agre. *Computation and Human Experience*. Cambridge University Press, 1997.

[6] Meta AI. Llama 3.2 3b: Efficient multilingual large language model. `https://huggingface.co/meta-llama/Llama-3.2-3B`, 2024. Accessed: 2024-12-10.

[7] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa. A systematic literature review of user trust in ai-enabled systems: An hci perspective. *International Journal of Human–Computer Interaction*, 40(5):1251–1266, Mar 2024.

[8] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[9] A. Bussone, S. Stumpf, and D. O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169, Dallas, TX, USA, Oct. 2015. IEEE.

[10] C. J. Cai, J. Jongejan, and J. Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 258–262, Marina del Ray California, Mar. 2019. ACM.

[11] Hyesun Choung, Prabu David, and Arun Ross. Trust in ai and its role in the acceptance of ai technologies. *arXiv preprint arXiv:2203.12687*, 2022. Accessed: 2024-12-10.

[12] Y.-N. Chuang et al. Efficient xai techniques: A taxonomic survey. *arXiv*, 2023. Accessed: May 09, 2024.

[13] Deloitte. Ai: Transforming the future of banking. Technical report, Deloitte, n.d. Accessed: 2024-12-10.

[14] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *Stat*, 1050:2, 2017.

[15] R. De Brito Duarte, F. Correia, P. Arriaga, and A. Paiva. Ai trust: Can explainable ai enhance warranted trust? *Human Behavior and Emerging Technologies*, 2023.

[16] U. Ehsan and M. O. Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. *arXiv*, 2020. Accessed: May 09, 2024.

[17] James Fogarty, Jacob O. Wobbrock, and Scott E. Hudson. Personalized human-computer interaction at scale: Challenges and opportunities. *Human-Computer Interaction*, 32(2):81–125, 2017.

[18] Leo A. Goodman. Snowball sampling. *Annals of Mathematical Statistics*, 32(1):148–170, 1961.

[19] Ben Green and Salomé Viljoen. Algorithmic realism: Expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 19–31. ACM, 2020.

[20] Ben Green and Salomé Viljoen. Algorithmic realism: Expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 2020)*, pages 19–31. ACM, 2020. Accessed: 2024-12-10.

[21] R. Hoffman, S. Mueller, G. Klein, and J. Litman. Measuring trust in the xai context. PsyArXiv Preprints, 2021.

[22] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. *arXiv*, 2021. Accessed: May 09, 2024.

[23] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 379–390, Marina del Ray, California, 2019. ACM.

[24] T. Le, T. Miller, R. Singh, and L. Sonenberg. Explaining model confidence using counterfactuals. *AAAI*, 37(10):11856–11864, Jun 2023.

[25] Zachary C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[26] D. Harrison McKnight, Michelle Carter, and Paul Clay. Trust in technology: Development of constructs and measures. In *Proceedings of the 15th Americas Conference on Information Systems*, 2009. Accessed: 2024-12-10.

[27] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[28] John Murawski. Mortgage providers look to ai to process home loans faster. *Wall Street Journal*, March 2019. Retrieved 16-September-2020.

[29] M. Nauta et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, 55(13s):1–42, Dec. 2023.

[30] Dat Nguyen, Angelica Martinez, and Jessica Horacio. Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, 55:2453–2486, 2021. Accessed: 2024-12-10.

[31] T. Nguyen, A. Canossa, and J. Zhu. How human-centered explainable ai interface are designed and evaluated: A systematic survey. *arXiv*, 2024. Accessed: May 09, 2024.

[32] United Nations Department of Economic and Social Affairs. Technological change and sustainable development. Technical report, United Nations, 2019. Accessed: 2024-12-10.

[33] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability, 2018.

[34] DJ Pangburn. Schools are using software to help pick who gets in. what could go wrong? *Fast Company*, May 2019. Retrieved 16-September-2020.

[35] Y. Rong et al. Towards human-centered explainable ai: A survey of user studies for model explanations. *arXiv*, 2023. Accessed: May 07, 2024.

[36] M. Roszel, R. Norvill, J. Hilger, and R. State. Know your model (kym): Increasing trust in ai and machine learning. *arXiv*, 2021. Accessed: May 09, 2024.

[37] Interactive Schools. 50 million users: How long does it take tech to reach this milestone?, n.d. Accessed: 2024-12-10.

[38] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 59–68, New York, NY, USA, 2019. ACM.

[39] A. Smith-Renner, V. Kumar, J. Boyd-Graber, K. Seppi, and L. Findlater. Digging into user control: perceptions of adherence and instability in transparent models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 519–530, Cagliari, Italy, 2020. ACM.

[40] Latanya Sweeney. Discrimination in online ad delivery: A multidisciplinary inquiry. *Communications of the ACM*, 56(5):44–54, 2013.

[41] ToolTester. Chatgpt statistics: Adoption, usage, and trends, n.d. Accessed: 2024-12-10.

[42] Author Unknown. Artificial intelligence governance: Balancing innovation and regulation. *ERA Forum*, 21(4):613–624, 2020. Accessed: 2024-12-10.

[43] H. Ding Y. Li, S. Wang and H. Chen. Large language models in finance: A survey. *Human Behavior and Emerging Technologies*, 2024.

[44] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? 2017.

[45] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 2018.

Appendix C: Interview Protocol

Introduction

Thank you for joining this interview. We'll be discussing your experiences with the AI system you interacted with, particularly focusing on how trust and explainability played a role in your perception of the system. There are no right or wrong answers, and your input will be invaluable.

Broad Opening Question

To begin, could you tell me about your overall experience with the AI system? What stood out to you, positively or negatively?

1. Trust in Information Quality

Accuracy and Completeness

How did the accuracy of the AI's explanations affect your trust in its outputs?

Can you recall any situations where the explanation felt incomplete? How did that impact your perception?

Consistency and Continuity

Did you notice if similar inputs yielded similar explanations? How did that affect your trust in the AI system?

Were there moments where inconsistent explanations stood out to you? What was your reaction?

Contrastivity

Did the system address "why not?" or "what if?" scenarios effectively? Could you share examples where this worked well or fell short?

2. Presentation of Explanations

Compactness and Composition

How did the length or style of the explanations influence your trust?

Did you find concise explanations more helpful, or did detailed ones make you feel more confident?

Confidence Indicators

How did the presence (or absence) of confidence levels or probabilities affect your trust?

What kind of information would make confidence indicators more meaningful to you?

3. User-Centric Design

Relevance and Context

Were the explanations relevant to your needs or the task you were trying to accomplish? Can you give an example?

What do you think the AI could do to make its explanations more useful for you?

Coherence and Controllability

How well did the explanations align with your existing knowledge or expectations? Could you describe moments where this worked well or didn't?

Did you feel in control of the AI's explanation process? What features would enhance that sense of control?

Conclusion

Wrap up:

Is there anything else about your experience with the AI system that significantly affected your trust or understanding that we haven't discussed?