Exploring the effects of explainability on model trust

Allen Daniel Sunny

November 15, 2024

1 Abstract

The advent of GPT4 marks the emergence of large-scale AI applications, signaling a new era in technological adoption. These systems, exemplified by ChatGPT, are beginning to permeate various subsectors across various industries including law[47], medicine[19] and business [28], potentially altering the social landscape and highlighting concerns about stability and trust. [2],[34] This study explores the crucial relationship between explainability and user trust in AI systems. By contrasting different types of explainability, including a novel approach involving interactive counterfactual explanations, we aim to determine how these methods influence trust. Our findings suggest that interaction not only enhances trust but that counterfactual explanations significantly boost the confidence users have in AI systems. This paper builds on previous [10] research, which has shown that traditional feature-based explanations often fall short, by investigating alternative forms of explainability that may better satisfy the demands for user understanding and trustworthiness in AI.

2 Introduction

The rapid expansion of large language models like ChatGPT marks a significant evolution in artificial intelligence, profoundly influencing society in diverse ways. Often, the most

marginalized groups are disproportionately impacted, experiencing the greatest risks and minimal benefits from AI advancements. [46] This stark inequality highlights the critical need for models that are not just technically advanced but are also understandable and trustworthy for all users, especially those most susceptible to negative implications.

Instances abound where the deployment of AI has raised concerns, such as in criminal justice [15] and healthcare sectors [29], underscoring the urgency for more responsible AI development. Despite these challenges, the momentum behind AI research continues unabated. Historical evidence suggests that attempting to decelerate the advancement of AI technologies or any new technology is not feasible.[42]

A promising approach involves reframing AI development through a human-centered perspective, succinctly defined as Human-Centered Explainable AI [11], providing individuals with the necessary tools to effectively engage with complex machine learning models. Current explainability methods often adopt a one-size-fits-all strategy, utilizing tools like feature importance scores and textual counterfactuals which may not suffice for all users. True comprehension of an AI system lies at the intersection of what users find trustworthy and what experts believe is reliable.

Research indicates that interactivity and tailored explanations, such as counterfactual reasoning, enhance understandability,[39] [36] thereby fostering trust. These methods allow users to see how different inputs affect outputs and to explore hypothetical alternatives, offering a more transparent view of AI decision-making processes.

This paper expands on existing research by examining various explainability techniques to establish a clear relationship between how AI systems explain their decisions and the trust users have in them. By comparing different methods, we aim to identify the most effective strategies that enhance both the transparency and reliability of AI outputs.

Furthermore, this introduction sets the stage for a deeper examination into which explainability methods resonate most effectively with users. We explore the dynamics of user interaction with AI explanations, advancing our understanding of how to increase transparency in AI

to build trust.

3 Related Work

The field of Explainable Artificial Intelligence (XAI) has witnessed significant evolution, transitioning from developer-centric methods toward approaches that prioritize user engagement and understanding.[30] [13] Initial XAI efforts highlighted technical transparency, utilizing inherently interpretable models such as linear regressions and decision trees.[28] However, as model complexity has increased, so has the necessity for advanced post-hoc exploratory techniques. Zhu et al. have criticized the focus on developing new XAI algorithms at the expense of usability and practical interpretability for real users [31].

3.1 Types of Explanations

XAI explanations are typically categorized as either local or global:

- Local explanations are designed to clarify the decision-making process for individual instances, making them highly relevant to end users who prioritize understanding specific decisions over system-wide behaviors. [22] [9]

- Global explanations provide a comprehensive view of the model's overall behavior, but they may not address individual user concerns as effectively as local explanations. Given our focus on individual decision-making, local explanations are more pertinent to this study. [35]

Explanations can also be generated as either pre-hoc or post-hoc:

- Pre-hoc explanations are integrated into the model during its design and are less complex but often less accurate than black box models.

- Post-hoc explanations are derived from fully trained models. These are crucial for explaining 'black box' AI systems, particularly at a local level, where they provide the most value to end-users. [35] [7]

3.2 Categories of Post-hoc Explanations

Research typically categorizes post-hoc explanations into several types, including:

- Feature Importance Explanations (e.g., LIME, SHAP): These methods highlight how each input feature impacts the model's output, offering a straightforward approach to understanding model decisions.[35]

- Simplification methods: These explanations show how minor modifications to input data could lead to different outcomes, aligning with human tendencies to engage in counterfactual thinking. [25]

- Example-based explanations: Interactive adjustments by users that demonstrate how varying inputs alter outcomes, thereby enhancing user understanding and trust. [6]

3.3 Interactive Methods and Games

Recent studies have started to explore the use of interactive methods and games as dynamic tools to assess and amplify the relationship between explainability and trust. [13] These methods are believed to not only bolster user engagement but also deepen insights into how different types of explanations affect user perception and confidence in AI systems. [48]

The consensus is that explainable AI models play a pivotal role in understanding AI system decisions and enhancing user confidence and trust in these systems. Previous research has indeed explored and substantiated the assumption that AI explainability positively impacts trust.[27] [26] [31] Additionally, evidence suggests that counterfactual explanations, due to their natural contrastive attributes aligning with human causal reasoning, offer a

valuable means of explaining models. [25] [23] And that counterfactual explanations mixed with interactive explanations can broaden the scope of explainability significantly.

Factors such as interactions, system communication, and the system's mental model are pivotal in decision-making tasks. However, to our current knowledge, the impact of incorporating explanations into decision-making scenarios on trust in AI remains uncharted.

4 Research Questions

Based on the comprehensive review of the literature on AI trust and explainability discussed in the previous section, we have formulated the following research questions to guide our investigation into the relationship between trust and explainability in AI systems:

- **Research Question 1 (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?
- **Research Question 2 (RQ2)** What elements of these explanations are most influential in building or undermining users' trust during their interactions with this AI system?

These questions aim to empirically test the effects of different types of AI explanations on user trust, further contributing to our understanding of how explainability can be effectively implemented to foster greater acceptance and reliance on AI technologies.

5 Methods

We have developed a web-based game designed to test our hypotheses regarding AI explainability in the context of loan approval processes. The game leverages a sophisticated AI assistant that provides recommendations on loan approvals based on a set of defined criteria, which players can either accept or reject. This experimental approach utilizes the loan

approval game—a recognized method for evaluating AI explainability [33], incorporating both demographic and monetary information.

5.1 Game Development

The game is structured as a web-based application, presenting players with fifteen scenarios where decisions to approve or reject a loan must be made based on demographic data (gender, geographical region, race, age) and five key monetary attributes (total loan amount, repayment period, interest rate, credit score, and debt-to-income ratio). These attributes have been identified as critical factors in determining loan approvals. On-screen, players are shown all relevant information, along with the AI system's recommendation for accepting or rejecting each loan. After each decision, players receive feedback indicating whether their choice was favorable. Furthermore, the AI provides explanations for its recommendations, and if the model is interactive, players can request further clarifications through a text box before proceeding to the next scenario. The primary objective of the game is to minimize the approval of bad loans, with player performance quantified through a final score.

5.2 AI Development

To ensure robust decision-making, we employed a CatBoost tree-based classifier trained on a dataset from the UCI machine learning repository [49], encompassing both demographic and financial data. The dataset comprises 16 masked variables, including six continuous and ten categorical variables, with the target variable indicating loan approval or disapproval.

5.2.1 Good AI System

The good AI system was trained to exclude demographic data from its decision-making process, thereby eliminating potential biases. It achieved an accuracy of approximately 90% on the training set, ensuring reliable and ethical decision-making.

5.2.2 Bad AI System

In contrast, a flawed AI model was developed by randomizing 40% of the target variables in the dataset, leading to a reduction in accuracy to 65%. This model serves as a control to assess the impact of AI reliability on user trust and understanding, providing a comparative measure against the more reliable system.

6 Measuring Trust

Trust is a complex, multifaceted concept extensively studied across various disciplines. Establishing trust between humans already presents numerous challenges, but fostering trust between humans and artificial intelligence (AI) is essential for widespread AI adoption. According to research detailed in sources [1], [37], and [20], the factors influencing trust in AI can be categorized into human-based, context-based, and technology-based dimensions. Notably, context-based factors can even include inherent distrust of AI systems, such as the skepticism surrounding AI in weapons systems noted in [21]. Trust is not static but evolves continuously and dynamically, far beyond a one-time establishment [44].

Explainability is highlighted in sources [1] and [37] as a crucial component for building trust, with [1] distinguishing between transparency and interpretability—both of which are vital for fostering trust. Additional influential factors include an AI system's perceived empathy [4], [40], as well as non-technical aspects like privacy, fairness, and accountability. From a technical perspective, characteristics such as safety, accuracy, and robustness are highly valued, often viewed by users as essential guarantees of trust. Further research, including sources [45] and [32], emphasizes a preference for accuracy over explainability when establishing trust. Moreover, certain cybersecurity studies [43] suggest that increased robustness in models also enhances trust, [43] also indicates that trust fundamentally stems from the reliability and safety of the system.

Given the breadth of literature, there is a significant established link between trust and

explainability in AI systems. This study will delve deeper into this relationship. While the primary focus will be on explainability, other critical factors such as fairness, robustness, and transparency can also be considered to ensure a comprehensive understanding of trust dynamics in AI systems.

6.1 Trust Scales

There have been some measures that historically measured trust. [14] detailed one of the earliest methods of looking at trust within an organization. [16] adapted trust scales from [5]. Below we have the scale we are using to measure trust. We have marked the main idea of the question in bold.

The scale consists of five items that participants respond to on a Likert-type scale ranging from strongly disagree to strongly agree:

6.1.1 Description of the Scales

- **Confidence:** "I am confident in the [tool]. I feel that it works well."
- **Predictability:** "The outputs of the [tool] are very predictable."
- **Reliability:** "The tool is very reliable. I can count on it to be correct all the time."
- **Accuracy:** "I feel safe that when I rely on the [tool] I will get the right answers."
- **Efficiency:** "The [tool] is efficient in that it works very quickly."
- **Distrust:** "I am wary of the [tool]." (Adapted from the Jian, et al. Scale and the Wang, et al. Scale)
- **Accuracy:** "The [tool] can perform the task better than a novice human user." (Adapted from the Schaefer Scale)
- **Preference for Decision Making:** "I like using the system for decision making."

A copy of scales used is attached in the appendix.

7 Measuring Explainability

The concept of explainability within artificial intelligence (AI) models has evolved significantly, driven by diverse measurement approaches. Scholarly contributions have elaborated and quantified explainability through various methodologies. A comprehensive summarization of these methods is presented in [30], which summarizes twelve categories encapsulated under the acronym COP-12, including Correctness, Completeness, Consistency, and others. Additionally, [8] delineates explainability into two overarching categories: User Explanation Satisfaction and System Explanation Satisfaction, simplifying the multifaceted nature of explainability assessment. Further, [41] categorizes explainability according to different philosophical and scientific paradigms, offering a multidisciplinary perspective on explainability evaluation. Another segmentation [3] introduces explainability into five distinct categories, including Understandability and Comprehensibility. These explorations reveal a rich tapestry of methodologies for measuring model explainability, underscoring varied perspectives and criteria employed by researchers. Given the aim of our study—to assess whether an increase in the quality of explanations enhances user trust—we focus on metrics that directly relate to user perception of these explanations.

7.1 Explainability Scales

The development of scales to measure the explainability of AI models is still evolving, with initial efforts informed by human-robot interaction studies [38]. A more refined scale is presented in [12], marking a significant advancement in the measurement of AI explainability. Further enhancements are discussed in [18]; a detailed scale based on the SUS scale [24] illustrates ongoing improvements in this area. For our research, we employ the scale published in [17], which builds upon the SUS scale and encompasses a broad spectrum of explainability

aspects previously discussed.

7.2 Description of the Scale

The scale consists of several items that participants respond to on a Likert-type scale ranging from strongly disagree to strongly agree:

- **Assesses Understanding:** "From the explanation, I know how the [software, algorithm, tool] works."
- **Measures Satisfaction:** "This explanation of how the [software, algorithm, tool] works is satisfying."
- **Evaluates Detail:**"This explanation of how the [software, algorithm, tool] works has sufficient detail."
- **Tests Completeness:**"This explanation of how the [software, algorithm, tool] works seems complete."
- **Checks Usability:**"This explanation of how the [software, algorithm, tool] works tells me how to use it."
- **Assesses Utility:**"This explanation of how the [software, algorithm, tool] works is useful to my goals."
- **Accuracy:**"This explanation show me how accurate the [software, algorithm, tool] will be."

A copy of the scale has been included in the appendix for reference.

8 Participants and Data Collection

Participants were recruited from a combination of college campuses and surrounding neighborhoods to ensure a diverse mix of ages and backgrounds. The recruitment within college

primarily targeted departments known for their technical curriculum, such as computer science and data science, to ensure a baseline familiarity with technology among participants. This approach aimed to provide a balanced cross-section of the population, reflecting varied levels of AI competency and age. Power analysis was carried out to determine the optimal number of participants

Participants were categorized into three age brackets and three levels of AI competency as follows:

Age Brackets: Individuals were divided into three groups:

1. Young (18-25 years)
2. Middle-aged (26-45 years)
3. Older (46 years and above)

AI Competency Levels: Based on responses to a preliminary survey assessing familiarity with and understanding of AI, participants were classified into:

1. Novice: Little to no prior experience or understanding of AI.
2. Intermediate: Moderate experience and understanding, likely from educational exposure or casual use.
3. Expert: Extensive experience and understanding, typically from professional use or advanced study in relevant fields.

8.1 Power Analysis

A power analysis was conducted to determine the required sample size to detect significant effects with an alpha level of 0.01. The analysis indicated that a minimum sample size of 150 participants would be needed to achieve 80% power to detect a medium effect size in the study design.

9 Experimental Design

The study employs a 3x3x2 mixed factorial design to systematically examine the effects of age, AI competency, and AI system quality on trust and user interaction outcomes. This design facilitates the assessment of both main effects and interaction effects among the three variables, providing a comprehensive analysis of how different factors influence user perceptions and behaviors with AI systems.

9.1 Factors Involved

9.1.1 Age Brackets (3 levels)

Young (18-25 years), Middle Aged(26 -45 years), Older(46 years and above)

9.1.2 AI Competency Levels (3 levels)

Novice, Intermediate and Expert

9.1.3 AI System Type (2 levels)

Good AI: Characterized by high reliability and accuracy and Bad AI: Characterized by lower reliability and accuracy.

Participants are randomly assigned to one of the nine combinations of age brackets and AI competency levels, ensuring a balanced distribution across all categories. Within these groups, each participant interacts with both types of AI systems (Good and Bad) in a counterbalanced order to control for any order effects and to ensure that each participant's experience with one type of AI does not influence their perceptions of the other.

9.2 Variable Definitions

9.2.1 Two Types of AI Systems

- Good AI (high accuracy)

- Bad AI (low accuracy)

9.2.2 Four Levels of Explainability

- No explanation
- Basic explanation (feature importance)
- Detailed explanation (contextual information)
- Interactive explanation (allows user queries)

9.2.3 Independent Variables

- **AI System Type (Within-Subjects):** Participants interact with both the Good AI and the Bad AI systems. The order of interaction is counterbalanced to control for order effects and fatigue.
- **Level of Explainability (Between-Subjects):** Each participant is randomly assigned to one of four explanation conditions and remains in that condition throughout the study.

9.2.4 Dependent Variable

- **Trust:** Measured using a standardized scale after interactions with each AI system. This captures shifts in trust based on the AI type and the level of explainability provided.

10 Procedure

Upon recruitment, participants are first briefed about the study's objectives and the nature of their involvement. They provide informed consent before proceeding with the tasks.

Participants are initially categorized into one of three age groups and one of three AI competency levels based on responses to a preliminary survey assessing their familiarity with and understanding of AI technologies. This categorization is crucial for ensuring a diverse representation in the study.

Participants are then randomly assigned to interact with two types of AI systems—'Good AI' and 'Bad AI'—in a counterbalanced order. This counterbalancing is essential to control for order effects, ensuring that the experience with one system does not influence the perception of the other. Each participant engages in a series of tasks designed to simulate real-world scenarios that require reliance on the AI system's decision-making capabilities. These scenarios are embedded in a web-based game interface, which serves both to present the tasks and to collect participants' responses seamlessly.

After each interaction with an AI system, participants complete a standardized survey designed to measure their trust and satisfaction with the AI's decisions. This feedback is critical for assessing the impact of AI explainability on user trust.

To ensure consistency across sessions, all interactions are conducted in a controlled environment, either in a lab setting or remotely, depending on the participant's location. The web-based game is accessible on various devices, accommodating different participants' preferences and ensuring broad accessibility. Each session is designed to last approximately one hour, with breaks included to mitigate fatigue.

The data from each interaction are automatically captured by the game interface and stored securely for later analysis. At the end of their session, participants are debriefed, giving them the opportunity to ask questions and provide feedback about their experience. This debriefing also serves to reiterate the confidentiality of their responses and the non-commercial purpose of the study.

Through this detailed and systematic procedure, we aim to gather rich data on how different demographic groups perceive and trust AI systems under varied conditions. The insights derived from this data will inform the design of more user-centered AI technologies.

A data flow diagram of the set up is attached in the appendix as Figure 8. Examples of the interfaces are also attached in the appendix.

11 Qualitative Findings and Discussion

In this section, we analyze the open-ended discussion conducted at the end of the interaction session. This discussion explored how users perceive explainability and which aspects of explainability contribute to building trust in the system. Responses were coded using the Cop-12 metrics, and we developed codes aligned with the broader themes within these metrics.

The main themes identified are: [1] Content, [2] Presentation, and [3] User-Focused Aspects. These themes encompass the Cop-12 sub-themes. However, since our focus is primarily on user-centered factors, we concentrate on the Cop-12 themes most relevant to this perspective.

Given the granularity of the sub-themes, our analysis emphasizes the impact on the broader themes without delving deeply into sub-theme coding. When providing participant responses as examples, participants are anonymized and referred to as P1 (Participant 1), P2 (Participant 2), and so on. It is worth noting that certain sub-themes may be coded across multiple responses.

11.0.1 Content

This dimension gauges the content that has been generated by the system. It has dimensions such as correctness (How faithful is the explanation to the black box) , completeness (How much of the behavior is described in the explanation) , consistency(Same input same explanation) , continuity (Similar inputs should be explained in similar ways), constravity(Answers Why/Why Not questions) and co-variate complexity (Human understandable concepts about feature interaction). With respect to consistency , (P1) noted that the sys-

tem gave "...For very similar incomes, it gave different outcomes." (P1) also noted that the consistency "...For very similar incomes, it gave different outcomes." was lacking. Both metrics indicated a loss of trust.

11.0.2 Presentation

This dimension talks about how exactly the explanations have been presented to the user. It has dimensions such as Compactness(length of the explanation) , Composition (Presentation) and Confidence(Probability information of the system). (P1) talked about how there was no confidence metric associated with the system "...I wanna know how confident the machine itself is with itself." and how that lowered trust in the system. Regarding Compactness, (P1) also talked about how detailed the explanations should have been. "...If it was more detailed, I would trust the decision more."

11.0.3 User Focused

This dimension gauges how the explanations are viewed by the user. This involves things such as the context (How useful is it to the user?) , coherence (Reasonable sounding) and the controllability of the the explanations (Can the user influence the decisions). (P1) talked about how "..I had control over the different parameters I could send into it, but not what it was giving out." This highlights that user control is important in building trust in the system.

12 Conclusion

References

[1] S. Afroogh, A. Akbari, E. Malone, M. Kargar, and H. Alambeigi. Trust in ai: Progress, challenges, and future directions. *arXiv*, 2024. Accessed: May 09, 2024.

[2] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv*, 2016. Accessed: May 07, 2024.

[3] A. B. Arrieta et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *arXiv*, 2019. Accessed: May 09, 2024.

[4] M. Ashoori and J. D. Weisz. In ai we trust? factors that influence trustworthiness of ai-infused decision-making processes. Available online at http://arxiv.org/abs/1912.02675, Dec 2019.

[5] Béatrice Cahour and Jean-François Forzy. Does projection into use improve trust and exploration? an example with a cruise control system. *Safety Science*, 47(9):1260–1270, 2009. ffhal-00471270f.

[6] 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.

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

[8] X. Cui, J. M. Lee, and J. P.-A. Hsieh. An integrative 3c evaluation framework for explainable artificial intelligence. Provide additional details here if available.

[9] M. Du, N. Liu, and X. Hu. Techniques for interpretable machine learning. *arXiv*, 2019. Accessed: May 08, 2024.

[10] 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.

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

[12] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl. Automated rationale generation: A technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274, Marina del Ray, California, 2019. ACM.

[13] L. B. Fulton, J. Y. Lee, Q. Wang, Z. Yuan, J. Hammer, and A. Perer. Getting playful with explainable ai: Games with a purpose to improve human understanding of ai. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, Honolulu HI USA, Apr. 2020. ACM.

[14] N. Gillespie. Measuring trust in organizational contexts: An overview of survey-based measures. In F. Lyon, G. Möllering, and M. Saunders, editors, *Handbook of Research Methods on Trust*, pages 175–188. Edward Elgar Publishing, 2012.

[15] M. R. Haque, D. Saxena, K. Weathington, J. Chudzik, and S. Guha. Are we asking the right questions?: Designing for community stakeholders' interactions with ai in policing. In *Conference Name (if known)*. ACM, 2024.

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

[17] R. Hoffman, S. Mueller, G. Klein, and J. Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5, 2023.

[18] A. Holzinger, A. Carrington, and H. Müller. Measuring the quality of explanations: The system causability scale (scs). comparing human and machine explanations. *Künstl Intell*, 34(2):193–198, Jun 2020.

[19] Y. Huang, K. Tang, and M. Chen. A comprehensive survey on evaluating large language model applications in the medical industry. *arXiv*, 2024. Accessed: May 07, 2024.

[20] 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.

[21] J. Johnson. Artificial intelligence, drone swarming and escalation risks in future warfare. *The RUSI Journal*, 165(2):26–36, Feb 2020.

[22] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5686–5697, San Jose, California USA, 2016. ACM.

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

[24] Jeffrey R. Lewis. The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590, 2018.

[25] Y. Li, M. Xu, X. Miao, S. Zhou, and T. Qian. Prompting large language models for counterfactual generation: An empirical study. *arXiv*, 2024. Accessed: May 07, 2024.

[26] A. Malhi, S. Knapic, and K. Främling. Explainable agents for less bias in human-agent decision making. In D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling, editors, *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, volume 12175 of *Lecture Notes in Computer Science*, pages 129–146, Cham, 2020. Springer International Publishing.

[27] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 397–407, Marina del Ray California, Mar. 2019. ACM.

[28] R. Movva, S. Balachandar, K. Peng, G. Agostini, N. Garg, and E. Pierson. Topics, authors, and institutions in large language model research: Trends from 17k arxiv papers. *arXiv*, 2024. Accessed: May 07, 2024.

[29] A. Muley, P. Muzumdar, G. Kurian, and G. P. Basyal. Risk of ai in healthcare: A comprehensive literature review and study framework. *AJMAH*, 21(10):276–291, Aug 2023.

[30] 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.

[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] M. Nourani, S. Kabir, S. Mohseni, and E. D. Ragan. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proc AAAI Conf Hum Comput Crowdsourc*, volume 7, pages 97–105, 2019.

[33] M. Olckers and T. Walsh. Incentives to offer algorithmic recourse. *arXiv*, 2023. Accessed: May 09, 2024.

[34] I. D. Raji and R. Dobbe. Concrete problems in ai safety, revisited. *arXiv*, 2023. Accessed: May 07, 2024.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin. '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*, pages 1135–1144, San Francisco, California USA, 2016. ACM.

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

[37] 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.

[38] Kristin Schaefer. *The Perception and Measurement of Human-robot Trust*. Ph.d. thesis, Name of University, 2013. Available online at Electronic Theses and Dissertations.

[39] J. Schoeffer, N. Kuehl, and Y. Machowski. 'there is not enough information': On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1616–1628, Jun 2022.

[40] N. N. Sharan and D. M. Romano. The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, 6(8), Aug 2020.

[41] F. Sovrano and F. Vitali. An objective metric for explainable ai: How and why to estimate the degree of explainability. *Knowledge-Based Systems*, 278:110866, Oct 2023.

[42] I. Struckman. Examining experts' motivations for signing the 'pause letter'.

[43] M. Taddeo, T. McCutcheon, and L. Floridi. Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nat Mach Intell*, 1(12):557–560, 2019.

[44] S. Thiebes, S. Lins, and A. Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31(2):447–464, Jun 2021.

[45] N. Wang, D. v Pynadath, and S. G. Hill. Building trust in a human-robot team with automatically generated explanations, 2015. 12 pages, Available: files/5941/Wang%20et%20al.%20-%202015%20-%20Building%20Trust%20in%20a%20Human-Robot%20Team%20with%20Automati.pdf.

[46] J. Whittlestone and S. Clarke. *AI Challenges for Society and Ethics*, chapter 3 (if applicable). Oxford University Press, 2022.

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

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

[49] I-Cheng Yeh. Default of credit card clients. UCI Machine Learning Repository, 2016.

A Appendix Figures

From the explanation, I know how the [software, algorithm, tool] works.
This explanation of how the [software, algorithm, tool] works is **satisfying**.
This explanation of how the [software, algorithm, tool] works has **sufficient detail**.
This explanation of how the [software, algorithm, tool] works seems **complete**.
This explanation of how the [software, algorithm, tool] works **tells me how to use** it.
This explanation of how the [software, algorithm, tool] works is **useful to my goals**.
This explanation of the [software, algorithm, tool] shows me how **accurate** the [software, algorithm, tool] is.

Figure 1: Explainability Scale

1. I am confident in the [tool]. I feel that it works well.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

2. The outputs of the [tool] are very predictable.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

3. The tool is very reliable. I can count on it to be correct all the time.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

4. I feel safe that when I rely on the [tool] I will get the right answers.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

5. The [tool] is efficient in that it works very quickly.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

6. I am wary of the [tool]. (adopted from the Jian, et al. Scale and the Wang, et al. Scale)

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

7. The [tool] can perform the task better than a novice human user. (adopted from the Schaefer Scale)

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

8. I like using the system for decision making.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Figure 2: Trust Scale

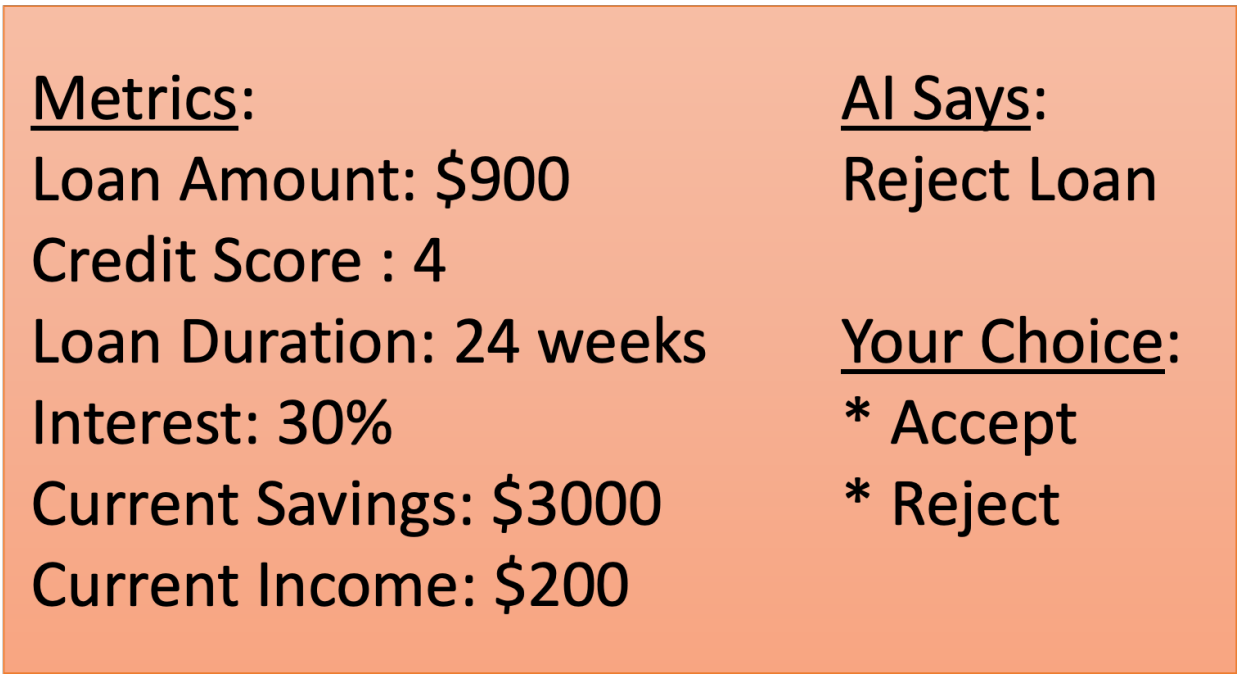

Figure 3: Example Interface

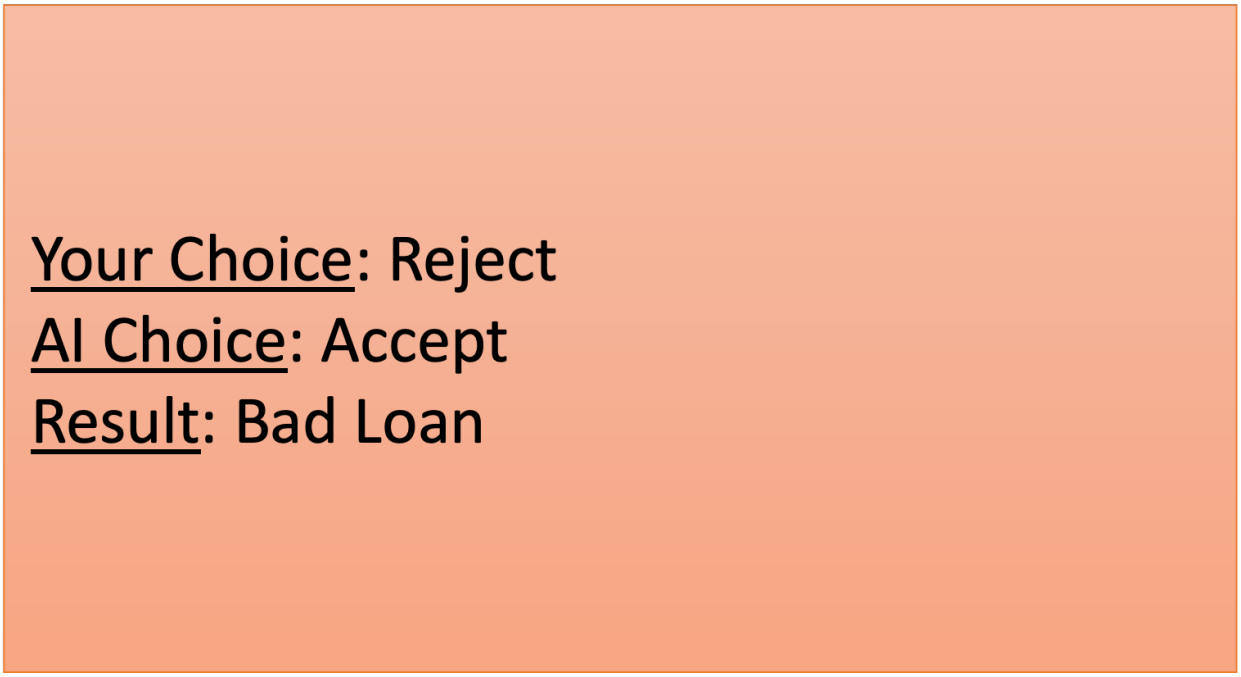

Figure 4: Example Interface

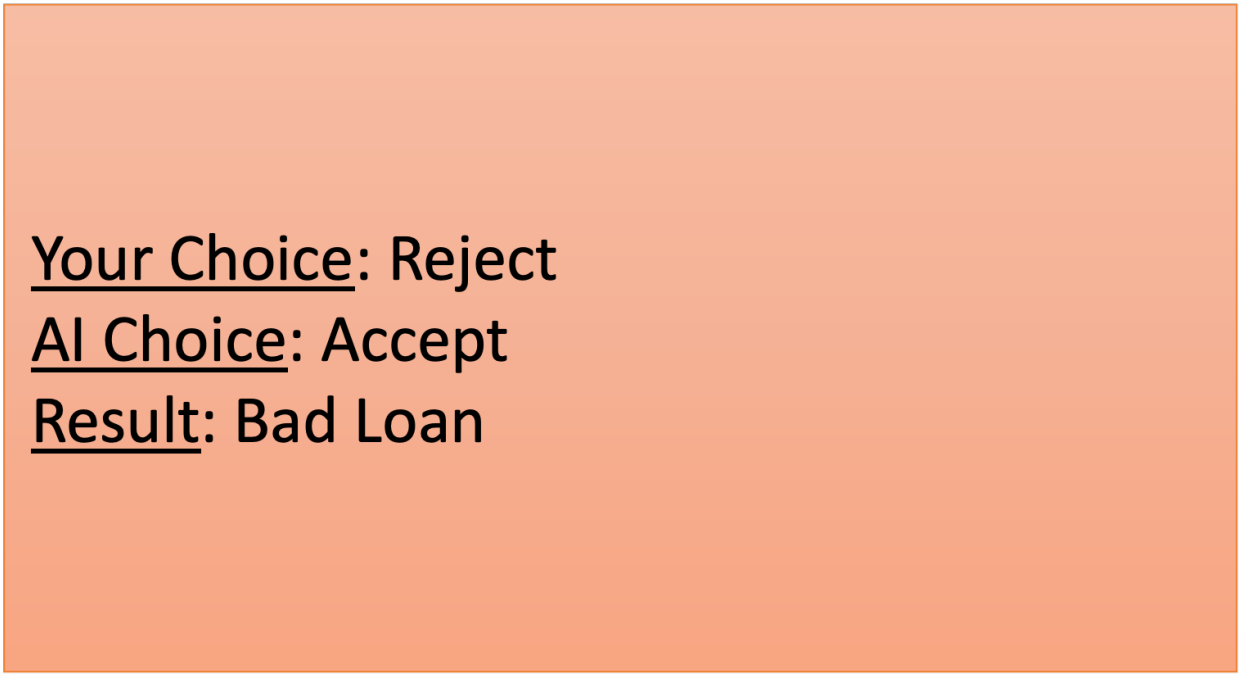

Figure 5: Example Interface

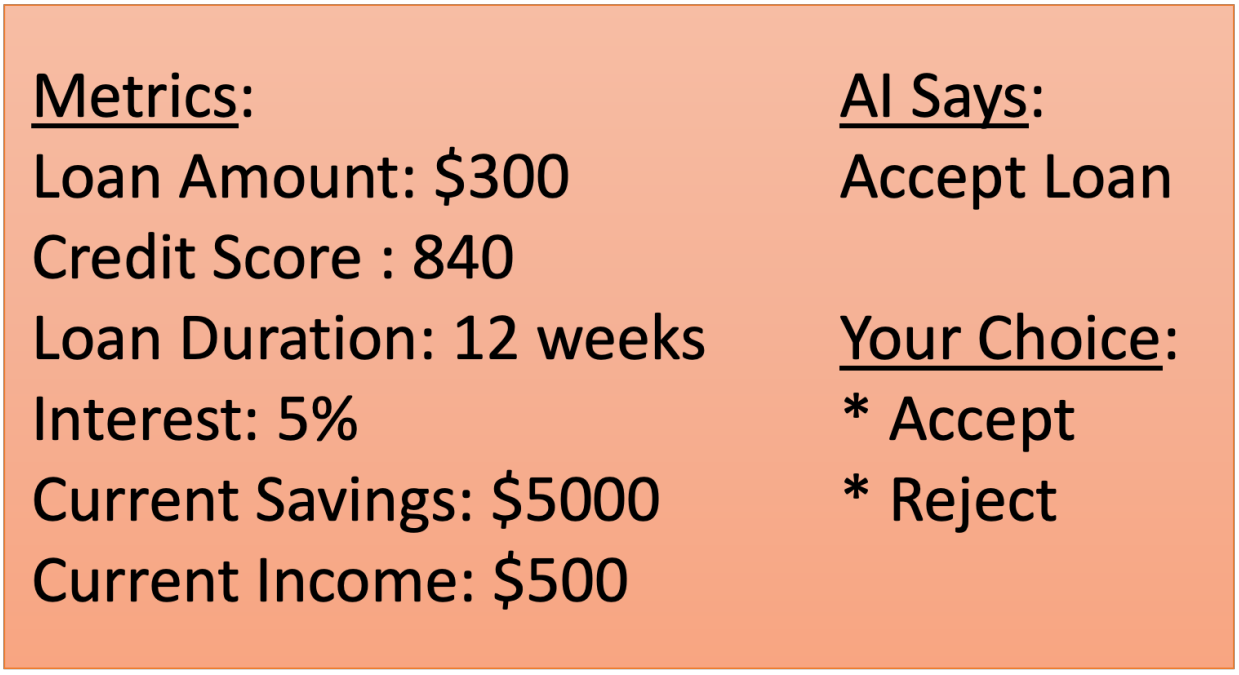

Figure 6: Example Interface

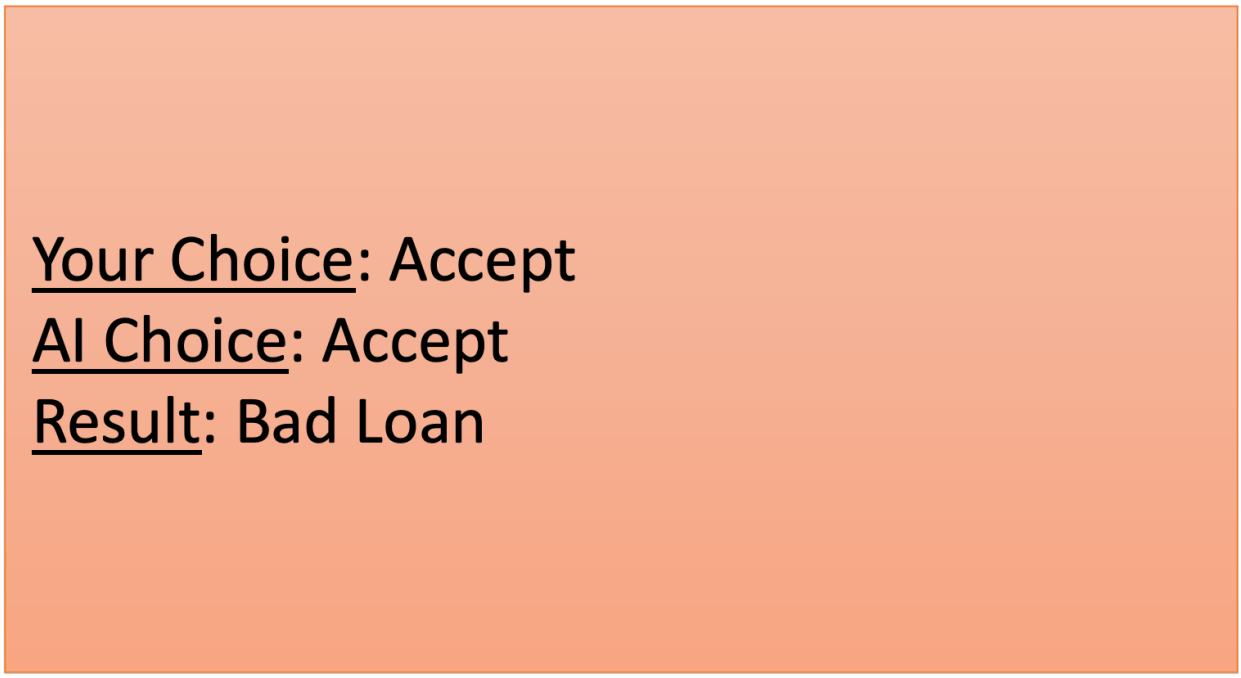

Figure 7: Example Interface

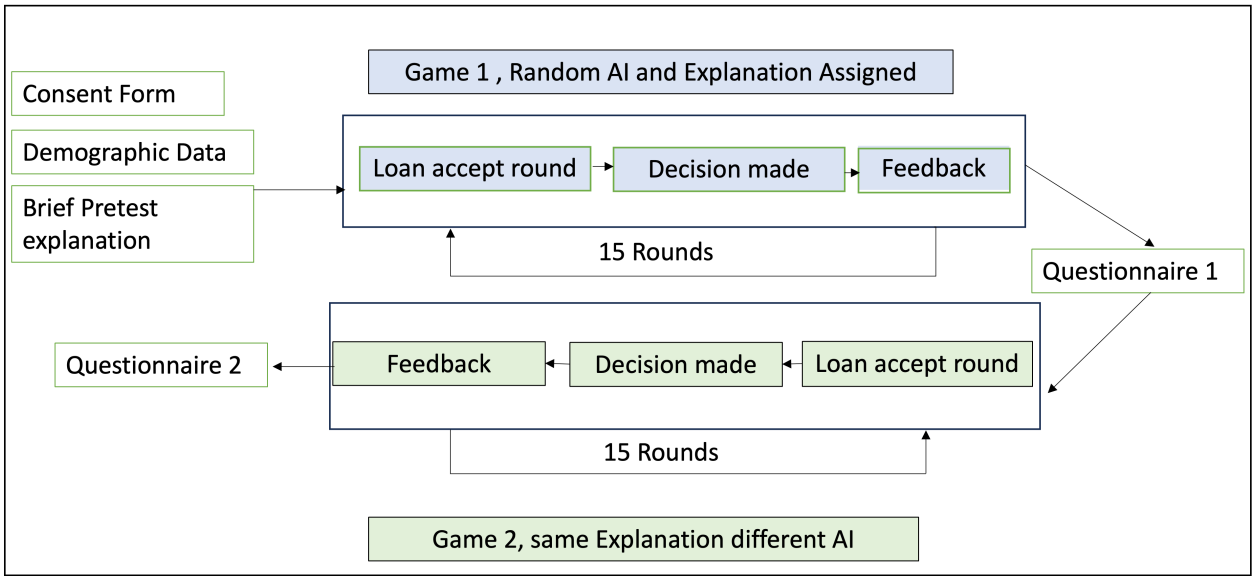

Figure 8: Procedure Diagram