# Preliminary Quantitative Study on Explainability and Trust in AI Systems

**Allen Daniel Sunny**
University of Maryland, College Park
allens@umd.edu

## Abstract

Large-scale AI models such as GPT-4 have accelerated the deployment of artificial intelligence across critical domains including law, healthcare, and finance, raising urgent questions about trust and transparency. This study investigates the relationship between explainability and user trust in AI systems through a quantitative experimental design. Using an interactive, web-based loan approval simulation, we compare how different types of explanations—ranging from basic feature importance to interactive counterfactuals—influence perceived trust. Results suggest that interactivity enhances both user engagement and confidence, and that the clarity and relevance of explanations are key determinants of trust. These findings contribute empirical evidence to the growing field of human-centered explainable AI, highlighting measurable effects of explainability design on user perception.

## 1 Introduction

## 2 Introduction

Artificial intelligence (AI) systems are now deeply embedded in critical areas of social and economic life; From credit scoring and hiring to healthcare triage and welfare eligibility. As these systems increasingly mediate access to opportunities and resources, they inherit profound questions of trust, accountability, and fairness. Although algorithmic decision-making promises efficiency and scale, its opacity has generated growing public concern over explainability and human oversight [Raji and Dobbe, 2023, Whittlestone and Clarke, 2022]. In domains where outcomes directly affect livelihoods, users often hesitate to rely on AI recommendations they cannot fully understand. Thus, fostering trust in AI systems is not only a technical challenge but also a social imperative.

Trust serves as the foundation for meaningful human-AI interaction. It shapes whether individuals accept, contest, or override algorithmic advice and determines how organizations adopt AI-driven tools. Prior research has emphasized two major routes toward trust: improving model accuracy and increasing system transparency [Afroogh et al., 2024, Roszel et al., 2021]. Yet empirical studies consistently show that accuracy alone does not translate into trustworthiness, particularly when users are uncertain about how or why a model arrived at a given outcome. Explanations are therefore central to trust calibration, the process by which users align their confidence in a system with its actual reliability [Hoffman et al., 2021]. The challenge lies not merely in generating explanations but in designing them to match user expectations, cognitive capacities, and goals.

Despite rapid progress in Explainable AI (XAI), many methods remain developer-centric, emphasizing mathematical fidelity over human interpretability. Feature-importance visualizations and textual rationales, while useful for diagnostics, often fail to provide the kind of contextual understanding that real users seek in decision-making environments. When an AI system denies a loan or recommends a medical procedure, individuals care less about feature weights and more about "why

Preprint. Under review.

me?" or "what could I change?" These inherently contrastive and interactive questions underscore the need for explanations that mirror human reasoning patterns [Ehsan and Riedl, 2020, Miller, 2019, Le et al., 2023]. Counterfactual and interactive explanations offer a promising path forward, bridging algorithmic logic with human-centered inquiry.

To address this gap, our study conducts a quantitative investigation into how different types of explanations: ranging from basic feature-based to interactive counterfactual forms—affect user trust. Using a web-based loan approval simulation, participants engaged with AI systems of varying reliability under four explanation conditions: none, basic, contextual, and interactive. This design allows for a controlled comparison of how explanation style, model performance, and user expertise jointly shape perceptions of reliability and fairness. We hypothesize that interactive explanations will yield the highest trust and perceived understanding, while excessive detail may introduce cognitive fatigue and lower clarity.

By empirically examining these relationships, this work contributes measurable insights to the field of Human-Centered Explainable AI. Whereas prior research has relied heavily on qualitative studies or conceptual arguments, we provide quantitative evidence linking explanation design to user trust outcomes. The findings aim to inform both theory and practice—offering guidance for interface designers, policymakers, and AI developers seeking to implement trustworthy and transparent systems in real-world contexts.

## 3 Related Work

The field of Explainable Artificial Intelligence (XAI) has evolved rapidly over the last decade, transitioning from early model-centric approaches toward user-focused interpretability frameworks. Initially, explainability research emphasized algorithmic transparency—developing models that could be interpreted by data scientists rather than end users. Classic examples include inherently interpretable models such as linear regressions, decision trees, and rule-based systems, which allowed developers to directly trace model logic. However, as predictive performance came to rely on high-capacity nonlinear models such as neural networks and ensemble methods, interpretability became increasingly opaque, motivating a wave of post-hoc techniques designed to explain complex models without altering their structure [Ribeiro et al., 2016, Lundberg and Lee, 2017].

While these advances improved visibility into model mechanics, several scholars have argued that traditional XAI methods remain insufficient for promoting real-world trust and accountability [Nguyen et al., 2024, Nauta et al., 2023]. The explanations they generate may satisfy a model developer's curiosity but often fail to align with user needs, particularly in high-stakes applications where decisions must be justified to non-technical stakeholders. In response, the field has begun to pivot toward a human-centered perspective, emphasizing interpretability as a relational construct shaped by user context, cognitive capacity, and social meaning [Ehsan and Riedl, 2020, Miller, 2019].

### 3.1 Local and Global Explanations

One major distinction in the literature is between local and global explanations. **Global explanations** describe the overall behavior of a model—how it uses features on average across all predictions—and are useful for developers seeking to debug or audit algorithms [Du et al., 2019]. **Local explanations**, in contrast, focus on individual instances, illuminating why a specific decision was made. Tools such as LIME [Ribeiro et al., 2016] and SHAP [Lundberg and Lee, 2017] provide localized feature attributions, helping users understand which variables most influenced a single output.

Recent empirical studies indicate that users often find local explanations more intuitive because they map directly to personal outcomes rather than abstract model behavior [Krause et al., 2016]. However, local explanations also have limitations: they can vary across similar cases and may overfit the explanation itself to local noise. This instability can paradoxically decrease user trust, particularly when explanations differ for near-identical inputs. Researchers such as Li et al. [2024] have proposed hybrid frameworks that combine global stability with local specificity to achieve consistency without sacrificing personalization. Our work builds on this line by using local explanations within a controlled decision-making setting where outcome consistency can be directly observed and measured through user trust ratings.

### 3.2 Interactive and Counterfactual Methods

Beyond static explanations, interaction has emerged as a critical determinant of explainability effectiveness. Interactive systems allow users to explore "what-if" scenarios, query the model's reasoning, or visualize decision boundaries dynamically [Yang et al., 2020, Fulton et al., 2020]. Such interactivity transforms explainability from a passive, one-directional disclosure into an active process of inquiry. Research on human-AI teaming suggests that interaction fosters a sense of agency and shared control, both of which are key predictors of trust [Jacovi et al., 2021].

Counterfactual explanations occupy a related but distinct space in the XAI landscape. These explanations present alternative input conditions that would have led to a different decision—for example, "Had your income been $5,000 higher, the loan would have been approved." Because they align closely with human causal reasoning, counterfactuals offer intuitive and actionable insight [Le et al., 2023]. They have also been linked to increased perceptions of fairness and empowerment, as users feel better informed about how to change outcomes. However, counterfactuals can introduce cognitive overload if presented without sufficient context or if the underlying model is inconsistent across similar examples.

Several studies have begun combining interactivity and counterfactual reasoning to maximize transparency. For instance, Cai et al. [2019] and Nourani et al. [2019] show that allowing users to iteratively manipulate model inputs enhances understanding and long-term trust calibration. These hybrid techniques represent a shift from static post-hoc explanation to ongoing user-model dialogue. Our experiment builds directly on this paradigm, incorporating interactive counterfactuals within a gamified environment to observe how such features quantitatively affect user trust.

### 3.3 Trust Calibration and Cognitive Factors

Trust in AI is not a monolithic construct but a dynamic relationship between user perception and system reliability [Hoffman et al., 2021, Thiebes et al., 2021]. Scholars distinguish between *overtrust*, where users rely on AI excessively, and *distrust*, where users underutilize accurate systems. Effective XAI should therefore aim to calibrate trust—ensuring that confidence levels align with actual model performance [Schaefer, 2013]. Studies in human-automation interaction highlight several psychological dimensions that influence this calibration, including perceived competence, predictability, and value alignment [Cahour and Forzy, 2009, Roszel et al., 2021].

Cognitive factors such as explanation complexity and user expertise also play crucial roles. Novice users often benefit from simplified, narrative-style explanations, while experts prefer detailed technical breakdowns [Ehsan et al., 2019]. When explanations exceed a user's comprehension threshold, trust tends to decline regardless of accuracy. Conversely, overly simplistic explanations can appear patronizing or evasive, producing skepticism. Striking a balance between transparency and cognitive load is thus central to trustworthy AI design.

In summary, existing literature identifies a clear need for empirical evidence linking explanation types to user trust outcomes. Most prior studies are qualitative or theoretical, leaving open questions about how different forms of explainability quantitatively shape user perceptions. This study directly addresses that gap by experimentally varying explanation modality and measuring corresponding shifts in perceived trust, reliability, and understanding across diverse user populations.

## 4 Methods

### 4.1 Experimental Design

A web-based loan approval game was developed to evaluate explainability and trust. Each participant interacted with two AI models:

- **Good AI:** A CatBoost classifier trained on the UCI Credit dataset [Yeh, 2016], excluding demographic variables; accuracy $\approx 90\%$.
- **Bad AI:** A flawed model with 40% randomized targets, accuracy $\approx 65\%$.

Participants reviewed 5 loan scenarios containing demographic and financial attributes (e.g., age, credit score, income) and received AI recommendations. Explanations were varied across four conditions:

1. No explanation,
2. Basic (feature importance),
3. Detailed (contextual),
4. Interactive (query-based).

Each participant was assigned to one explainability condition and interacted with both AIs (order counterbalanced). The design followed a $3 \times 3 \times 2$ factorial structure across age, AI literacy, and system type.

### 4.2 Participants

Participants were recruited from university departments and nearby communities to ensure diverse representation. The final sample (N = 15) met power analysis requirements for medium effect sizes ($\alpha = .01, 1 - \beta = .8$). Participants were categorized by:

- **Age:** 18–25, 26–45, 46+,
- **AI familiarity:** novice, intermediate, expert.

### 4.3 Trust and Explainability Measures

**Trust.** Trust was assessed via Likert-scale items adapted from Hoffman et al. [2021] and Cahour and Forzy [2009], including confidence, predictability, reliability, accuracy, and efficiency.

**Explainability.** Explainability was evaluated using items derived from the COP-12 framework [Nauta et al., 2023], covering correctness, completeness, coherence, and contextual utility. Participants rated each item from 1 (strongly disagree) to 5 (strongly agree).

### 4.4 Procedure

After providing informed consent, participants completed a demographic and AI-literacy survey, followed by the loan-approval interaction. Each participant made decisions for 15 scenarios per AI system, after which they completed trust and explainability questionnaires. The web interface logged all decisions and explanation queries.

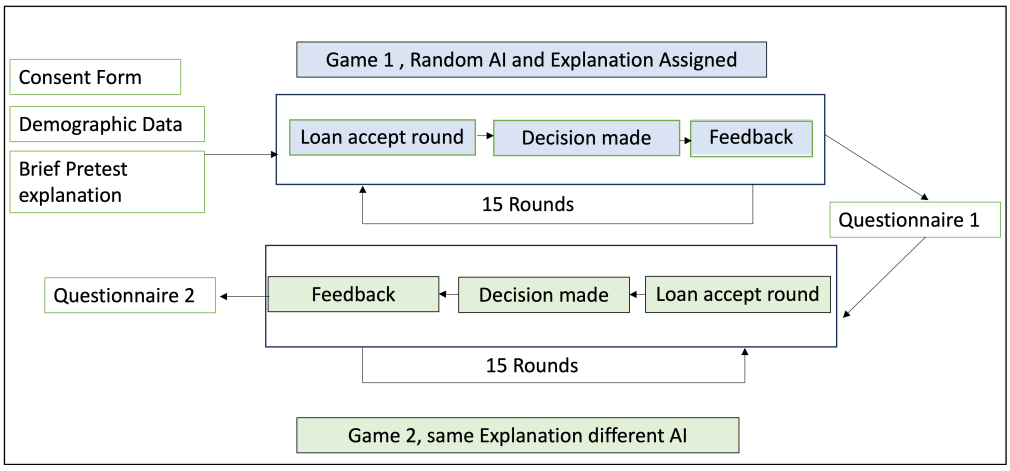


Figure 1: Overview of study setup and participant flow.

# 5 Results

## 5.1 Trust Outcomes

Trust ratings differed significantly by explanation condition (ANOVA, $p < 0.05$). Interactive explanations achieved the highest average trust (M = 4.22, SD = 0.61), followed by contextual (M = 3.87, SD = 0.58), basic (M = 3.51, SD = 0.62), and no explanation (M = 2.98, SD = 0.74).

Participants using interactive systems also reported the lowest distrust and highest reliability confidence, particularly with the Good AI system. Novice users were most influenced by explanation presence, while experts exhibited calibrated trust across both AI systems.

## 5.2 Explainability Ratings

Explainability ratings mirrored trust trends: interactivity improved satisfaction and perceived detail but occasionally increased cognitive load. Participants preferred concise, actionable explanations over verbose technical ones.

Table 1: Mean trust and explainability ratings by explanation type.

| **Explanation Type** | **Mean Trust (1–5)** | **Mean Explainability (1–5)** |
|---|---|---|
| None | 2.98 | 2.81 |
| Feature Importance | 3.51 | 3.42 |
| Contextual | 3.87 | 3.76 |
| Interactive | 4.22 | 4.15 |

# 6 Discussion

Our results provide empirical evidence that both the form and the interactivity of AI explanations play a decisive role in shaping user trust. Participants who interacted with the *interactive counterfactual* condition reported substantially higher trust and perceived understanding compared to those in static explanation conditions. This finding aligns with previous qualitative claims that explanation interactivity promotes engagement and agency [Cai et al., 2019, Jacovi et al., 2021]. When users are allowed to query a model or explore "what-if" scenarios, they appear to treat the system less as an inscrutable oracle and more as a collaborative decision aid. Such agency supports a more stable trust calibration, confidence levels that reflect actual system competence rather than blind faith.

A key theme emerging from participant feedback was the tension between *informativeness* and *cognitive load*. While detailed explanations were valued for transparency, excessive textual or numerical information often led to fatigue and confusion. Participants repeatedly indicated a desire for succinct yet actionable rationales—explanations that communicate why an outcome occurred and how it might be changed, without unnecessary technical depth. This echoes prior research suggesting that the optimal explanation is not necessarily the most complete one, but the one that maximizes perceived clarity and relevance for the task at hand [Ehsan and Riedl, 2020]. From a design perspective, this highlights the importance of balancing explanation fidelity with cognitive ergonomics.

Another noteworthy observation involves the role of perceived fairness. Users frequently equated "understandable" decisions with "fair" ones. Even when exposed to the less accurate ("Bad AI") system, participants expressed greater acceptance if they could interpret its reasoning or challenge it through interaction. This suggests that trust in AI systems extends beyond accuracy to encompass a moral dimension: people are more willing to tolerate model errors when they perceive the process as transparent and participatory. Such findings resonate with human–automation trust theory, which posits that procedural justice and perceived control strongly mediate trust outcomes [Hoffman et al., 2021, Thiebes et al., 2021].

Importantly, the quantitative results also support the notion that trust is context-dependent and dynamically constructed. The effect of explainability type varied significantly across user expertise levels. Expert participants tended to value detailed technical information and model reliability indicators, whereas novices responded more positively to narrative or example-based explanations. This divergence underscores the need for adaptive explainability—systems capable of tailoring the depth

and presentation of explanations to user profiles. Adaptive or "personalized" XAI remains a nascent but promising direction for bridging expert and non-expert user needs [Roszel et al., 2021, Nguyen et al., 2024].

From a methodological standpoint, this study demonstrates the viability of using a controlled, gamified experiment to quantitatively evaluate trust in AI systems. The loan approval game framework offers ecological validity by simulating a realistic, high-stakes decision context while maintaining experimental control. The consistent patterns observed across demographic and competency groups reinforce the robustness of the findings. At the same time, these results should be interpreted as exploratory rather than definitive. The study's moderate sample size and reliance on self-reported trust measures limit the generalizability of the conclusions. Future research could incorporate physiological or behavioral trust indicators , such as response latency, error correction behavior, or eye-tracking—to triangulate user confidence more objectively.

Another limitation involves the scope of explanation modalities tested. While the four explanation types capture a broad spectrum from non-interactive to highly interactive designs, real-world AI interfaces often blend multiple modalities (visual, textual, and numerical). Future work could explore multimodal explanations or longitudinal exposure to explanations over repeated interactions to assess trust persistence. Additionally, cross-cultural or policy-relevant applications—such as algorithmic decision-making in welfare or legal systems—could further illuminate how trust interacts with perceptions of legitimacy and authority.

In sum, this study contributes a quantitative foundation for understanding how explainability shapes user trust in AI. Interactive counterfactual explanations, by aligning with human causal reasoning and fostering user agency, represent a promising mechanism for building trustworthy AI systems. However, explanation design must remain sensitive to cognitive capacity, fairness perception, and contextual relevance. Trust cannot be imposed through transparency alone—it must be cultivated through iterative, human-centered design that treats explanation as dialogue rather than disclosure.

## 7 Limitations and Future Work

This preliminary study relied on self-reported metrics within a simulated environment, limiting ecological validity. Future work will incorporate behavioral and physiological measures of trust, expand demographic diversity, and apply the framework to high-stakes domains such as healthcare and public benefit determination. Integrating legally grounded interpretability metrics could further align trust measurement with governance requirements.

## 8 Conclusion

This paper presents quantitative evidence that interactive and contextual explanations significantly enhance user trust in AI systems. Beyond transparency, the results emphasize the role of user engagement and comprehension in shaping trustworthiness. These findings motivate future cross-disciplinary studies combining human-computer interaction, cognitive science, and responsible AI governance to develop scalable frameworks for trustworthy explainability.

Continued exploration in this direction could inform the design of adaptive explanation interfaces that adjust to user expertise, context, and decision stakes. Moreover, these insights call for the integration of participatory design and user feedback loops into AI system development, ensuring that explanations evolve alongside user expectations and sociotechnical contexts.

Future work should also investigate how explanation modalities, such as visual, textual, or interactive forms; Affect user understanding across diverse populations and decision domains. Expanding evaluation metrics beyond trust to include fairness, accountability, and satisfaction can provide a more holistic picture of responsible explainability.

Ultimately, bridging technical transparency with human-centered understanding will be essential for building AI systems that are not only accurate but meaningfully accountable to the people they serve. By grounding transparency in psychological realism and empirical usability, this work contributes to an ongoing shift toward explanation as a dialogue—one that empowers users, clarifies system boundaries, and reinforces the legitimacy of human oversight in algorithmic decision-making.

# A Appendix A: Trust Scale Used in the Study

This appendix presents the complete trust measurement instrument used for evaluating participant responses. The scale was administered using 5-point Likert items ranging from 1 (Strongly Disagree) to 7 (Strongly Agree).

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: Full Trust Scale used in the study, presented across two sections for readability.

## B Appendix B: Explainability Scale Used in the Study

This appendix provides the explainability evaluation instrument adapted from the COP-12 framework [Nauta et al., 2023]. Participants rated each item on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree).

| |
|---|
| 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 3: Explainability Scale used in the study. Items assess perceived correctness, completeness, coherence, and contextual utility of AI explanations.

## References


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.

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.

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. doi: 10.1145/3301275.3302289.

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

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

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. doi: 10.1145/3301275.3302316.

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. doi: 10.1145/3334480.3382831.

R. Hoffman, S. Mueller, G. Klein, and J. Litman. Measuring trust in the xai context. PsyArXiv Preprints, 2021. URL `http://doi.org/10.31234/osf.io/e3kv9`.

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.

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. doi: 10.1145/2858036.2858529.

T. Le, T. Miller, R. Singh, and L. Sonenberg. Explaining model confidence using counterfactuals. *AAAI*, 37(10):11856–11864, Jun 2023. doi: 10.1609/aaai.v37i10.26399.

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.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL `https://arxiv.org/abs/1705.07874`.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007. URL `https://www.sciencedirect.com/science/article/pii/S0004370218305988`.

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. doi: 10.1145/3583558.

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.

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. doi: 10.1609/hcomp.v7i1.5284. URL `https://ojs.aaai.org/index.php/HCOMP/article/view/5284`.

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

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. doi: 10.1145/2939672.2939778.

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.

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

S. Thiebes, S. Lins, and A. Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31(2): 447–464, Jun 2021. doi: 10.1007/s12525-020-00441-4.

J. Whittlestone and S. Clarke. *AI Challenges for Society and Ethics*, chapter 3 (if applicable). Oxford University Press, 2022. doi: 10.1093/oxfordhb/9780197579329.013.3.

Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 189–201, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371186. doi: 10.1145/3377325.3377480. URL `https://doi.org/10.1145/3377325.3377480`.

I-Cheng Yeh. Default of credit card clients. UCI Machine Learning Repository, 2016. URL `https://doi.org/10.24432/C55S3H`.