Designing for Trust in AI: A Quantitative Study of Explainability at the Interface Level

Jeremy Bregulla
Department of IT and Technology
IU International University of Applied Sciences
Erfurt, Germany
email: jeremybregulla@gmail.com

Marius Schönberger
Department of IT and Technology
IU International University of Applied Sciences
Erfurt, Germany
email: marius.schoenberger@iu.org

Sibylle Kunz
Department of IT and Technology
IU International University of Applied Sciences
Erfurt, Germany
email: sibylle.kunz@iu.org

Abstract— **This work investigates how User-Centered Design principles can be applied to enhance explainability in Large Language Model interfaces. While technical advances in explainable Artificial Intelligence have focused predominantly on algorithmic transparency, much less attention has been given to user needs and perceptions. A quantitative study with 52 participants evaluated user perceptions of usability and explainability across several Large Language Model interface mockups. Results show that user-centered explainability design contributes positively to user comprehension of the system and its behavior, supporting the assumption that interface-level explainability can make complex Artificial Intelligence systems more understandable. Furthermore, the study found a positive relationship between explainability and usability, suggesting that well-designed explainability features not only improve understanding but also increase perceived usability. The findings highlight the need to integrate explainability into early stages of interface design and suggest design features for developing intelligible, usable, and trustworthy Artificial Intelligence.**

Keywords–Artificial Intelligence; Human-Computer Interaction; User-Centered Design; Trust; Explainability.

I. INTRODUCTION

Recent advances in Artificial Intelligence (AI), including autonomous driving [1], video based AI content detection [2] and methods for mitigating hallucinations in Large Language Model (LLM) [3] have intensified discussions around trust in AI. Unlike earlier technologies, AI systems solve complex problems, engage in natural-language interaction, and provide domain-specific solutions based on large knowledge bases. These capabilities challenge traditional interaction concepts and make multidisciplinary evaluation of the dynamics of Human-AI Interaction (HAI) increasingly important [4]. Trust in AI can be defined as the user's expectation that a system and its provider behave reliably and fulfill legal and ethical responsibilities throughout the interaction [5] [6]. Trust also involves the belief that the system will support user goals despite uncertainty and vulnerability [7].

Research on trust in automation provides an important foundation for trust in AI. Autonomous systems may be understood as forms of automation that select information and act toward goals with minimal human intervention, while humans retain supervisory authority [7] [8]. Preliminary research identified the automation conundrum: as automation becomes more reliable, operators may lose situational awareness and become less able to intervene effectively [9]. As AI systems become more autonomous, the risk of out-of-the-loop errors increases, leading to reduced user control. Conversely, establishing user control contributes to reliable task completion.

Users often develop distorted perceptions of AI capabilities. Some overestimate performance, resulting in automation bias, while others underestimate the risks of autonomous decision-making [10]. Initially, users approach AI with skepticism because limited prior experience makes system behavior less predictable [11]. Related to this is users' disproportionate loss of confidence in algorithmic performance when confronted with system errors [12]. This phenomenon, often referred to as algorithmic aversion, contrasts with a preference for human decision-makers, even in cases where human performance is substantially less accurate than that of the algorithm. Establishing trust in AI thus requires supporting users in forming an accurate understanding of system capabilities and limitations, particularly in situations where errors might otherwise undermine confidence.

Effective collaboration with autonomous systems requires calibrating user trust to actual system capabilities. Calibrated trust aims to prevent both overtrust and distrust, each of which can result in inappropriate reliance or disuse [7]. Although this challenge is well known in automation research, the increasing capabilities of AI systems are creating new challenges for interaction design. In this context, explainability plays an increasingly important role in guiding system design [10] [13]. Explainability is driven by two key characteristics of AI: (i) model complexity and (ii) non-deterministic outputs [14]. Many AI systems function as black boxes whose internal decision processes

are difficult for non-experts to understand [15]. In addition, unlike conventional software, AI systems (particularly LLMs) may produce substantially different outputs even in response to small prompt changes. Clarifying system capabilities and limitations can help users form more accurate mental models, support initial trust, and improve human-AI collaboration [16].

Leveraging explainability in user interaction with AI systems can help transform opaque black-box models into more transparent glass-box systems. Design strategies for calibrated trust include communicating past performance, exposing relevant algorithmic processes or intermediate results, simplifying operational logic, and clarifying the algorithm's purpose and intended scope [7] [9]. Effective AI integration requires alignment with users' needs and behaviors via the User-Centered Design (UCD) approach. UCD was developed to support the alignment of technological solutions with the needs of clients and end users [17]. Its methodology centers on integrating user feedback, often through usability testing. Applied to explainability, a UCD perspective examines how users experience AI interactions and how explainability measures contribute to usability.

As AI stakeholders differ in goals, knowledge, and usage contexts, explanations must be tailored accordingly [18]. Insights from the social sciences suggest that interfaces should support user reasoning by presenting explanations together with plausible alternatives [19]. Effective explanations should also communicate what happened, what data were used, and when anomalous or unexpected behavior occurs. In this way, AI systems do not merely justify outputs but support users in evaluating options, maintaining control, and engaging meaningfully with the system.

Claims about explainability and usability must be validated through close examination of user interactions. This study adopts an interdisciplinary perspective on HAI [18], focusing on LLM-based systems, and explores how grounding explainability in UCD can help transfer established design practices to emerging AI technologies. Rather than addressing model transparency or algorithmic optimization, it examines how interface-level design elicits, presents, and supports explanations that are meaningful to users. By investigating user preferences for explainability features, this study aims to generate actionable insights for designing AI systems that are understandable, usable, and potentially conducive to calibrated trust. This work is guided by two research questions:

RQ1: *To what extent does User-Centered Design for explainability shape user understanding of intelligent systems?*

RQ2: *How do explainability-oriented UI features, informed by design principles, affect user perception of system usability?*

For the quantitative evaluation, two hypotheses are derived. *H1* tests whether User Interface (UI) design based on a user-centered approach to explainability positively influences user understanding of the system. *H2* tests whether LLM-based interfaces with design elements that improve explainability achieve higher perceived usability.

H1: *UI design based on a user-centered approach to explainability positively influences user understanding of the system.*

H2: *LLM-based interfaces that include design elements enhancing explainability will score higher in perceived usability.*

The paper is structured as follows: Section II describes the related work. Section III presents the research methodology. Section IV reports the results, followed by a conclusion and future work in Section V.

II. RELATED WORK

Related work in Explainable Artificial Intelligence (XAI) seeks to make complex AI systems and their outputs understandable to stakeholders, particularly in high-risk, domain-specific contexts [20]. It supports tasks, such as reviewing outputs, building trust, and diagnosing errors [21]. While XAI has advanced at the level of machine learning engineering, efforts to reduce user uncertainty and increase transparency have yet to fully reach the interface level [22].

The evaluation of XAI systems can follow several complementary approaches that may be applied independently or in combination [23] [24], including application-grounded, human-grounded, and functionally grounded evaluation, which assess interpretability through real-world tasks, controlled experiments, and proxy metrics, such as fidelity, sparsity, and stability. Current research in XAI shows critical gaps in substantiating claims about interpretability and usability. Suh et al. [25] conducted a large-scale analysis of 18,254 XAI papers and identified 253 that explicitly referenced human-related terms, of which only 128 included some form of human evaluation, revealing that merely 0.7% of XAI literature empirically validates claims of explainability.

Recently the focus has shifted towards human-centered XAI, which emphasizes explanations that are actionable and understandable for intended users, prioritizing user goals, contexts, and cognitive capacities rather than focusing solely on technical model interpretability [26] [18]. A broader, complementary perspective is Human-Centered AI (HCAI), which aims to shape AI system design around the needs of diverse stakeholders and to balance human autonomy with advanced computer automation [4] [27]. Within the context of HAI, explainability is not restricted to technical XAI methods for evaluation but involves a rigorous view of the dynamics of interaction.

Several frameworks have established design guidance for AI interfaces, including the widely cited general guidelines for Human-AI Interaction by Amershi et al. [14]. Some of these guidelines emphasize explainability through clear communication of system capabilities, behavior, performance, and outcomes. Similarly, Eiband et al. [22] propose a process-oriented framework for transparency design that supports more accurate user mental models of intelligent systems. Shneiderman [27] introduces the Prometheus Principles as part of the HCAI paradigm. When applied at the UI level, these principles aim to enable users to track system operations, understand the current system state, and relate outcomes to preceding actions, thereby strengthening perceived control and self-efficacy.

III. METHODOLOGY

To address the proposed research questions and hypotheses, a quantitative survey-based study was conducted. The following subsections describe the survey design, participants, and evaluation metrics.

A. *Survey Design*

The survey comprised three sections: an introductory demographic part, a guided evaluation of mockup designs with user ratings, and a final open-ended question for additional feedback. The demographic section collected data on age, academic background, occupational field, and prior experience with chat-based LLM systems. The main section of the survey was divided into three subsections, each introduced by brief instructions and accompanied by one or two mockups (see Figures 1 and 2). A vacation-planning scenario framed all examples in order to keep participants focused on the explainability features rather than changing task contexts. Participants were instructed to assume the role of a user interacting with the AI system.

Each section presented a specific UI interaction pattern in response to a user query within the vacation-planning scenario. Participants reviewed the presented interface elements and were asked to evaluate system usefulness and their satisfaction with the explanations. Each subsection used a seven-point scale for perceived usability, as recommended by Lewis et al. [28], and a five-point Likert scale, following Hoffman et al. [29], assessing explanation quality in terms of trust, clarity, and sufficiency of detail.

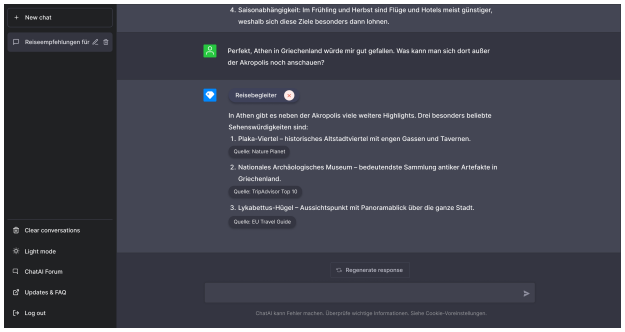

Figure 1. Mockup of evidence card (a) for the third subsection of the guided evaluation. The interface displayed a chat response with suggestions and clickable source buttons.

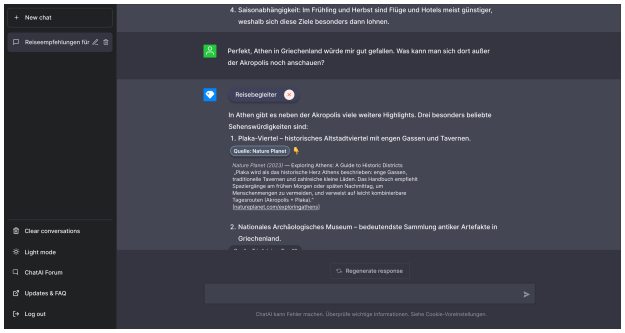

Figure 2. Mockup of evidence card (b) for the third subsection of the guided evaluation. The opened source tab displayed the referenced source supporting the suggestion.

The three mockups combined established and novel explainability interventions (see Table I). The general layout followed existing conversational AI systems, especially OpenAI's ChatGPT, while the visual design was deliberately simplified to reduce distraction and keep attention on the explainability elements. The mockup responses were based on a real-world query generated with ChatGPT 5. The first mockup introduced the role qualifier, which explicitly defined the AI system's function within the interaction based on the user's initial prompt. The subsequent mockups presented established explainability mechanisms: stepwise reasoning, which structured the AI's rationale sequentially, and evidence cards with hyperlinks, which clarified the data sources underlying the responses. The final section invited participants to provide qualitative feedback on HAI, the presented design features, and the survey itself.

TABLE I. OVERVIEW OF UI-DESIGNS FOR EXPLAINABILITY.

Feature	*Goal*	*ESS*
Role Qualifiers	Support user understanding of its intended behavior	Clarity
Stepwise reasoning	Prevent cognitive overload	Sufficiency of detail
Evidence Cards	Enhances output credibility	Trustworthiness

B. *Participants*

Participants were recruited through multiple channels. Employees of a small IT company and its parent company in B2C tech sales were invited to participate. The survey link was also shared via internal professional networks of a large German media company, specifically in groups related to technology, engineering, and AI. In addition, the survey was distributed through broader professional networks. The survey remained online for one month. A total of 55 responses were collected, of which 52 were retained after screening for completeness and response quality (n = 52).

Data were exported from Google Forms to Microsoft Excel for minor cleanup and then imported into VS Code for statistical analysis in a Jupyter Notebook. All collected data were processed in compliance with the GDPR. The demographic data indicated that nearly half of the participants (48.1%) held an academic degree. Most participants were long-term employees (82.7%), while only a small share were in early career stages (0–5 years) or identified as students or trainees. Prior exposure to LLMs was high: approximately 75% of participants rated themselves as experienced users (levels 3–5), whereas about 25% considered themselves novices (levels 1–2).

C. *Metrics*

To assess both usability and explainability constructs outlined in the hypothesis two standardized metrics designed for survey-based data collection were employed. Perceived usability was measured using the Usability Metric for User Experience (UMUX)-Lite, a shortened version of the original UMUX [28]. UMUX-Lite captures two core dimensions: functional adequacy ("This system's capabilities meet my requirements") and ease of use ("This system is easy to use"). Despite its brevity, the scale has demonstrated high reliability and strong correlation with System Usability

Score (SUS) scores, allowing conversion to the SUS scale.

Instead of using the standard two positively worded UMUX-Lite items, the survey adopted the SUS practice of alternating positive and negative wording to reduce response bias, as suggested by Brooks [30]. After translation and adaptation into German, the two items used were: "This system is useful for completing my tasks" and "This system is unnecessarily complex." The first item was slightly reformulated from "meeting requirements" to "useful for completing tasks" in order to better fit the survey scenario.

Explainability was assessed using selected items from the Explanation Satisfaction Scale (ESS) developed by Hoffman et al. [29]). Although originally designed for XAI systems, the ESS offers a structured way to evaluate how users perceive and experience explanations. It covers central dimensions, such as clarity, completeness, level of detail, usefulness, and trustworthiness.

In this study, subsets of ESS items were assigned to the different mockup sections according to the explainability features being evaluated [29]. The second ESS item ("This explanation of how the [software, algorithm, tool] works is satisfying") served as a global measure of explanation satisfaction across all mockups. Additional items were selected to match the explanatory goals of each design [29]: items 4, 7, and 8 for the first mockup; items 1 and 3 for the second; and items 5 and 6, merged into one question, together with item 9 for the third. This structure enabled the measurement of both overall explanation satisfaction and feature-specific explanatory goals.

IV. RESULTS

For quantitative testing, the survey assessed the two main variables usability and explainability using the aforementioned metrics. Descriptive statistics first provide an overview of both datasets. On the seven-point UMUX-Lite scale, overall usability was above the midpoint (mean 5.39, SD = 0.89) with Mockup 3 scoring highest (5.70) and Mockup 2 lowest (5.05). Variability was modest overall and highest for Mockup 1 (SD $\approx$ 1.06). For explainability (ESS), the overall mean was likewise above the midpoint (mean 3.29, SD = 0.71). Again, Mockup 3 scored highest (3.58) and Mockup 1 lowest (2.97), with moderate dispersion (SD $\approx$ 0.90).

Before inferential analysis, both datasets were tested for normality. Given the smaple size, the Shapiro–Wilk test and the Kolmogorov–Smirnov (KS) test were applied to the overall UMUX-Lite and ESS datasets. While the ESS data showed no significant deviation from normality, the UMUX-Lite did ($p < 0.001$). However, the non-significant KS-test results for both variables, together with visual inspection of histograms and kernel density estimates, were considered sufficient to proceed with parametric testing.

A. Usability Testing

The UMUX-Lite consists of two seven-point items. For evaluation, responses were transformed into SUS-equivalent scores (0–100 score) following Lewis et al. [28]. Prior to transformation, the negatively worded second item was reverse-coded using $x \mapsto 8 - x$ for $x \in \{1, \ldots, 7\}$. SUS-aligned usability scores were computed per participant and mockup, and the overall score was obtained by averaging across the three mockups. Following established SUS benchmarks and curved grading tables [28] [31], the average UMUX-Lite score across mockups was 70.40. This corresponds to a grade of C, typically around the 41st-59th percentile, indicating acceptable but improvable usability. At the mockup level, evidence cards achieved the highest SUS-aligned usability, followed by the role qualifier, while stepwise reasoning scored lowest (see Table II).

Table II. Usability results based on SUS benchmark.

Feature	*Mean*	*Grade*
Role Qualifier	70.61	C
Stepwise Reasoning	66.75	C
Evidence Cards	73.84	B–

B. Explainability Evaluation

The explainability analysis tested whether ESS scores exceeded the neutral midpoint. A one-sided t-test was conducted, testing $H_0 : \mu = 3.0$ against $H_1 : \mu > 3.0$ at $\alpha = 0.05$, resulting in $t(51) = 2.889$ and a one-sided $p = 0.0028$; with $n = 52$, $\bar{x} = 3.29$, and $sd = 0.71$.

The null hypothesis was therefore rejected, indicating that the UCD explainability design was associated with higher perceived explanation satisfaction, from which improved user comprehension can be inferred. The distribution of ESS scores (see Figure 3) supports this finding, with ratings concentrated above the midpoint (mean = 3.29, median = 3.28).

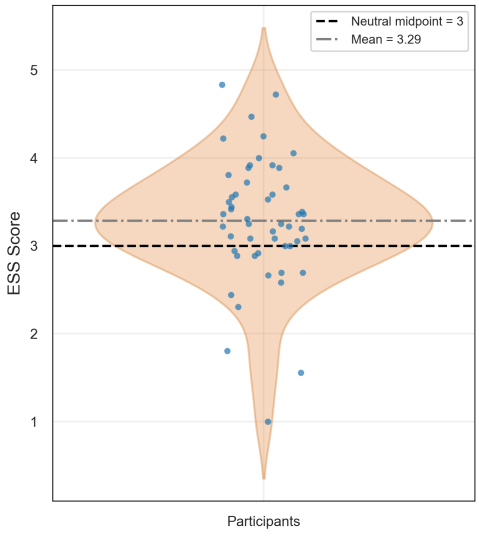

Figure 3. Explainability scores compared with neutral midpoint (3.0).

C. Bivariate Analysis

To examine the second hypothesis, a linear regression was conducted using overall ESS as predictor x and the transformed UMUX-Lite score as outcome y. The regression model is given by

$$\widehat{y} = 50.26 + 6.13\,x \qquad (1)$$

The relationship is moderate and statistically significant, with Pearson's correlation $r = 0.456$, coefficient of determination $r^2 = 0.208$, and a significant slope of $\hat{\beta}_1 = 6.13$

(p = 0.000688). These results present a positive linear relationship between explainability and perceived usability. Each one-point increase in ESS predicts an average 6.1-point increase on the usability scale. Explainability accounted for 20.8% of the variance in usability. Because ($p < 0.05$), the null hypothesis was rejected. The scatter plot with the regression line (see Figure 4) further illustrates this positive trend.

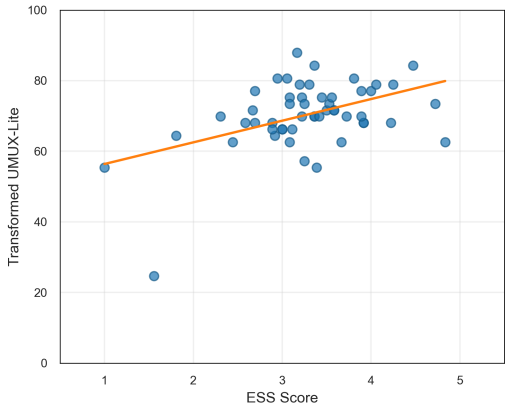

Figure 4. Linear regression on explainability and usability results.

V. DISCUSSION

This study examined two hypotheses related to UCD for explainability in LLM interfaces and its relationship to perceived usability. The first hypothesis addressed whether a user-centered approach to explainability supports user understanding of the system and its behavior. A one-sample, one-sided t-test on overall ESS provided support for this hypothesis (see 4.2). The findings indicate explanation satisfaction above the neutral midpoint, suggesting that UCD-based explainability can support user understanding of AI systems.

At the feature level, three design interventions were evaluated: (i) role qualifiers, targeting clarity and completeness; (ii) stepwise reasoning, addressing sufficiency of detail and usefulness; and (iii) evidence cards, focusing on trustworthiness. The global explanation item showed neutral to positive results across all three designs, with role qualifiers scoring = 2.97, stepwise reasoning = 3.31, and evidence cards = 3.58. Likewise, the proportion of favorable ratings (4 or 5) increased from 55.8% to 65.4% and 84.6%, respectively. These results indicate that evidence cards, which operationalize trustworthiness through visible sources, achieved the highest explanation satisfaction.

The usability results showed a similar pattern (see Table II). This suggests that evidence indicators may offer clearer usability benefits than social-role framing, while stepwise reasoning in its current form was perceived somewhat less favorably. The second hypothesis examined whether UCD-oriented explainability measurably affects perceived usability. A linear model with ESS predicting SUS-aligned UMUX-Lite scores revealed a moderate positive relationship, indicating that a one-point increase in explanation satisfaction corresponds to an average increase of 6.1 SUS-aligned points.

Overall, the findings suggest that user-centered explainability positively supports user comprehension and is positively associated with perceived usability. Beyond that, the results highlight that the effectiveness of explainability interventions depends on the specific explanatory goal they address. In particular, calibrated trust, implemented here through evidence cards, appears especially relevant for both explanation satisfaction and usability.

VI. CONCLUSION AND FUTURE WORK

This work quantitatively examined (i) the relationship between UCD for explainability and users' general understanding of an interface, and (ii) the relationship between explainability and perceived usability. To this end, a remote survey-based usability study was conducted using static mockups of a widely used chat-based LLM application. The survey included fifty-two participants, most of whom had an academic background and substantial prior experience with LLMs. The results indicate that interface-level explainability, when designed in a user-centered way, leads to general satisfaction with system explanations. They also reveal a positive relationship between explanation satisfaction and perceived usability, suggesting that users' evaluation of explainability is closely tied to their broader assessment of the interface.

Several limitations should be considered. First, the survey assumes that the implemented design features adequately reflect users' explainability needs. Although the mockups were based largely on established UI practices, prior feature-specific validation would strengthen this assumption. Second, the use of static mockups instead of interactive prototypes likely constrained the quality of participant feedback. Third, remote usability testing reduced control over participant behavior and testing conditions. The inclusion of an objective performance measure for usability could have further strengthened the evaluation. The recruitment strategy may have favored individuals who were more inclined to participate and complete the survey, introducing potential self-selection and nonresponse bias. Finally, it remains uncertain to what extent participants' ratings of explainability and usability referred specifically to the interface-level interventions rather than to the system outputs themselves, despite instructions to focus on the UI.

The observed relationship between explainability and usability has practical implications for UI design in AI systems. Explainability features that present explanations in a user-centered manner can improve comprehension by supporting clarity, sufficiency of detail, and trust in the provided explanations. These factors contribute to reduced cognitive load, increased confidence in explanation quality, and better understanding by limiting information to a level of sufficient detail. As a result, improved explanation satisfaction is associated with higher overall usability. These findings suggest a shift in AI development from a primarily technology-driven perspective toward a more user-centered approach, with the aim of enhancing the quality of everyday interactions with AI systems.

Future work should include larger and more diverse samples to enable the analysis of demographic and experience-related effects, while controlled experiments could better isolate the impact of interface design from properties of the model output. Additionally, the evaluation of explainability

features should incorporate interactive prototypes, mixed-method approaches, and objective performance-based usability measures. Advances in Human-Computer Interaction should therefore be accompanied by stronger efforts to realize explainability through interface design rather than focusing solely on model interpretability. Research that translates user needs for explainability into concrete user-centered features, evaluates them in realistic contexts, and iterates on them based on empirical evidence can advance the maturity of HAI.

REFERENCES

[1] S. Nordhoff and M. Hagenzieker, "'I will raise my hand and say "I over-trust Autopilot". I use it too liberally' – Drivers' reflections on their use of partial driving automation, trust, and perceived safety," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 107, pp. 1105–1124, 2024, doi:10.1016/j.trf.2024.09.021.

[2] Y. Zhou, E. Chen, M. Pisipati, A. Xiong, and S. Rajtmajer, "Effect of AI Performance, Risk Perception, and Trust on Human Dependence in Deepfake Detection AI System," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 7, art. CSCW407, 2025, doi:10.1145/3757588.

[3] A. Ryser, F. Allwein, and T. Schlippe, "Calibrated Trust in Dealing with LLM Hallucinations: A Qualitative Study," arXiv preprint arXiv:2512.09088, 2025.

[4] O. Ozmen Garibay et al., "Six Human-Centered Artificial Intelligence Grand Challenges," *International Journal of Human-Computer Interaction*, vol. 39, no. 3, pp. 391–437, 2023, doi:10.1080/10447318.2022.2153320.

[5] D. Shin and Y. J. Park, "Role of fairness, accountability, and transparency in algorithmic affordance," *Computers in Human Behavior*, vol. 98, pp. 277–284, 2019, doi:10.1016/j.chb.2019.04.019.

[6] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," *International Journal of Human-Computer Studies*, vol. 146, p. 102551, 2021, doi:10.1016/j.ijhcs.2020.102551.

[7] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004, doi:10.1518/hfes.46.1.50_30392.

[8] M. Fisher et al., "Towards a Framework for Certification of Reliable Autonomous Systems," arXiv:2001.09124, 2020.

[9] M. R. Endsley, "From Here to Autonomy: Lessons Learned From Human-Automation Research," *Human Factors*, vol. 59, no. 1, pp. 5–27, 2017, doi:10.1177/0018720816681350.

[10] Z. Atf and P. R. Lewis, "Human Centricity in the Relationship Between Explainability and Trust in AI," *IEEE Technology and Society Magazine*, vol. 42, no. 4, pp. 66–76, 2023, doi:10.1109/MTS.2023.3340238.

[11] G. Papagni, J. de Pagter, S. Zafari, M. Filzmoser, and S. T. Koeszegi, "Artificial agents' explainability to support trust: considerations on timing and context," *AI & Society*, vol. 38, no. 2, pp. 947–960, 2023, doi:10.1007/s00146-022-01462-7.

[12] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology: General*, vol. 144, no. 1, pp. 114–126, 2015, doi:10.1037/xge0000033.

[13] Z. Atf and P. R. Lewis, "Is Trust Correlated With Explainability in AI? A Meta-Analysis," *IEEE Transactions on Technology and Society*, pp. 1–8, 2025, doi:10.1109/TTS.2025.3558448.

[14] S. Amershi et al., "Guidelines for Human-AI Interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, Glasgow, Scotland, UK, pp. 1–13, 2019, doi:10.1145/3290605.3300233.

[15] R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, art. 93, pp. 1–42, 2019, doi:10.1145/3236009.

[16] P. Andras et al., "Trusting intelligent machines: Deepening trust within socio-technical systems," *IEEE Technology and Society Magazine*, vol. 37, no. 4, pp. 76–83, 2018, doi:10.1109/MTS.2018.2876107.

[17] B. Shneiderman et al., *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 6th ed. Boston, MA, USA: Pearson, 2016, ISBN: 978-1-292-15392-6.

[18] Q. V. Liao and K. R. Varshney, "Human-Centered Explainable AI (XAI): From Algorithms to User Experiences," arXiv preprint arXiv:2110.10790, 2022.

[19] T. Miller, "Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven decision support," arXiv preprint arXiv:2302.12389, 2023.

[20] S. Ali et al., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, p. 101805, 2023, doi:10.1016/j.inffus.2023.101805

[21] P. Raj, U. Köse, U. Sakthivel, S. Nagarajan, and V. S. Asirvadam, *Explainable Artificial Intelligence (XAI)*. London, UK: The Institution of Engineering and Technology, 2023, doi:10.1049/PBPC062E.

[22] M. Eiband et al., "Bringing transparency design into practice," in *Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18)*, Tokyo, Japan, pp. 211–223, 2018, doi:10.1145/3172944.3172961.

[23] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv preprint arXiv:1702.08608, 2017.

[24] S. Naveed, G. Stevens, and D. Robin-Kern, "An Overview of the Empirical Evaluation of Explainable AI (XAI): A Comprehensive Guideline for User-Centered Evaluation in XAI," *Applied Sciences*, vol. 14, no. 23, art. 11288, 2024, doi:10.3390/app142311288.

[25] A. Suh, I. Hurley, N. Smith, and H. C. Siu, "Fewer Than 1% of Explainable AI Papers Validate Explainability with Humans," arXiv preprint arXiv:2503.16507, 2025.

[26] M. Shajalal, "Towards Human-Centered Actionable Explainable AI-enabled Systems," Ph.D. dissertation, University of Siegen, 2025, doi:10.25819/ubsi/10665.

[27] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," arXiv preprint arXiv:2002.04087, 2020.

[28] J. R. Lewis, B. S. Utesch, and D. E. Maher, "UMUX-LITE: When there's no time for the SUS," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, Paris, France, pp. 2099–2102, 2013, doi:10.1145/2470654.2481287.

[29] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," arXiv preprint arXiv:1812.04608, 2019.

[30] J. Brooke, "SUS: A retrospective," *Journal of Usability Studies*, vol. 8, no. 2, pp. 29–40, 2013.

[31] J. Sauro and J. R. Lewis, *Quantifying the User Experience: Practical Statistics for User Research*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann, 2016, ISBN: 9780128025482.