ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(c) Jan – Feb, 2026

OPEN ACCESS Original Research Article

EXPLAINABILITY OVER ACCURACY: A HUMAN-CENTERED STUDY OF TRUST IN ARTIFICIAL INTELLIGENCE

* Kaushal Karthikeyan Nadar

Students, KSD's Model College, India .

Abstract:

As artificial intelligence becomes part of everyday decision-making, trust in these systems is no longer optional - it is essential. While most AI research focuses on improving accuracy, people often interact with systems that provide little to no explanation for their decisions. This study explores a simple but important question: do people trust AI systems that explain their decisions more than those that are highly accurate but opaque?

To examine this, we compare two simulated AI models. One model delivers highly accurate decisions without explanation, while the other provides clear, understandable explanations with slightly lower accuracy. Participants are presented with AI-generated decisions in a controlled scenario and are asked to evaluate their level of trust, perceived fairness, confidence, and willingness to rely on each system.

The results indicate that transparency plays a significant role in shaping user trust. Participants generally show a stronger preference for AI systems that offer explanations, even when they are informed that these systems may be marginally less accurate. Explanations help users feel more confident, involved, and assured that decisions are being made fairly.

These findings suggest that accuracy alone is not sufficient for building trustworthy AI. Instead, explainability should be treated as a core design principle, especially in applications where human judgment, accountability, and ethical concerns are critical.

Keywords: *Explainable Artificial Intelligence, Human Trust in AI, Transparency in AI Systems, Human-Centered AI, Decision Making Systems, User Perception of AI*

Introduction:

Artificial intelligence (AI) systems are increasingly being used to support or automate decision making in various domains such as education, healthcare, finance, and recruitment. From recommending candidates for admission to evaluating applications and predicting outcomes, AI-driven systems are no longer limited to technical environments but are actively interacting with human users. As the influence of these systems grows, the issue of trust becomes central to their acceptance and effective use.

Traditionally, the performance of AI systems has been evaluated primarily on the basis of accuracy. Higher accuracy is often assumed to indicate a better and more reliable system. However, many high-performing AI models operate as "black boxes", meaning that users are unable to understand how or why a particular decision has been made. This lack of transparency can lead to discomfort, skepticism, and reduce trust, especially in situations where decisions have significant personal or ethical consequences.

In response to this challenge, the concept of Explainable Artificial Intelligence (XAI) has gained attention. Explainable AI aims to make AI systems more transparent by providing users with understandable explanations for their decisions. Rather than focusing solely on predictive performance, XAI emphasizes interpretability, fairness and

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(c) Jan – Feb, 2026

OPEN ACCESS Original Research Article

accountability. For human users, explanations can help reduce uncertainty and create a sense of involvement in the decision making process.

Despite growing interest in explainability, an important question remains largely unexplored: do people value explainability more than accuracy when forming trust in AI systems? While technical research often prioritizes performance metrics, human trust may depend more on clarity and understanding than on marginal gains in accuracy. This gap highlights the need for human-centered studies that examine how users perceive and evaluate different types of AI systems.

This study investigates the relationship between explainability, accuracy and user trust by comparing two AI systems: one that provides highly accurate decisions without explanations and another that offers transparent explanations with slightly lower accuracy. Through a survey- based experiment conducted using a Google Form, the research aims to understand how explanations influence trust, perceived fairness, and willingness to rely on AI systems in decision making contexts.

Related Work:

Early research in artificial intelligence primarily focused on improving predictive performance, with accuracy treated as the main indicator of system quality. However, Doshi-Velez and Kim [1] argued that accuracy alone is insufficient for real-world deployment, particularly in high-stakes applications, and emphasized the need for interpretability and explainability to ensure trust and accountability.

Ribeiro et al. [2] made a significant contribution to explainable AI by introducing model-agnostic explanation techniques such as LIME. Their work demonstrated that providing local explanations improves user trust and understanding, even when the underlying model is complex. This showed that explanations help users assess whether AI decisions align with expectations and domain knowledge.

Lipton [3] critically examined the concept of interpretability and highlighted the risks of assuming that all explanations are equally meaningful. The study emphasized that poorly designed explanations may create a false sense of trust, underscoring the importance of clarity and relevance in explanation design.

From a human-centered perspective, Miller [4] showed that explanations aligned with human reasoning patterns enhance trust and acceptance. The study emphasized that simple and relevant explanations are more effective than overly technical ones.

Comprehensive surveys by Arrieta et al. [5] and Guidotti et al. [6] provided structured overviews of explainable AI methods and emphasized the role of transparency in fairness and trust. Additionally, the DARPA Explainable Artificial Intelligence (XAI) program described by Gunning and Aha [7] reflects growing institutional emphasis on explainable systems. Despite these contributions, limited empirical work directly compares high-accuracy black-box systems with explainable systems from a human-centered perspective. This gap motivates the present study, which evaluates user trust, perceived fairness, and willingness to rely on AI systems with differing levels of explainability.

Methodology:

A. Research Design

This study adopts a quantitative, survey-based research design to examine how explainability influences human trust in artificial intelligence systems. The experiment focuses on comparing user perceptions of two AI systems that differ primarily in their level of transparency. Data was collected using an online questionnaire to ensure uniform presentation of scenarios and standardized response collection.

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(c) Jan – Feb, 2026

OPEN ACCESS Original Research Article

B. Participants

Participants included individuals from diverse age groups and professional backgrounds. Responses were collected from students as well as working professionals, including acquaintances from the researcher's personal network. The survey was distributed online, and participation was entirely voluntary. To encourage honest responses, all participants remained anonymous, and no personally identifiable information was collected.

Participants reported varying levels of familiarity with artificial intelligence systems, allowing the study to capture a broader range of perspectives on trust and explainability in AI-driven decision-making.

C. Experimental Setup

The experiment involved two simulated AI systems presented to participants through a Google Form:

AI System A (Black-Box Model):

This system was described as having high accuracy but did not provide any explanation for its decisions.

AI System B (Explainable AI Model):

This system provided decisions along with clear, human-readable explanations but was described as having slightly lower accuracy compared to AI System A.

Both systems were evaluated using the same decision-making scenario to control for contextual variables. The only difference between the two systems was the presence or absence of explanations.

D. Scenario and Decision Context

Participants were presented with an admission-related decision scenario in an educational context. The problem description was kept brief and neutral, stating that an applicant was evaluated based on academic performance, entrance test results, and extracurricular activities. Detailed numerical information was intentionally omitted to prevent participants from forming independent judgments about the decision itself.

E. Data Collection Instrument

Data was collected using a structured Google Form consisting of four sections:

1. Participant information and consent
2. Evaluation of AI System A
3. Evaluation of AI System B
4. Direct comparison and final preference

For each AI system, participants rated their perceptions using five-point Likert scale questions measuring trust, perceived fairness, confidence, and willingness to rely on the system. Additional multiple-choice and short-answer questions were included to capture overall preferences and qualitative feedback.

F. Data Analysis

Responses were analyzed using descriptive statistical methods. Mean scores were calculated for each trust-related factor across both AI systems. Comparative analysis was performed to identify differences in participant preferences between the black-box and explainable AI models. Qualitative responses were reviewed to identify common themes that supported the quantitative findings.

G. Ethical Considerations

Ethical principles were followed throughout the study. Participation was voluntary, informed consent was obtained at the beginning of the survey, and all responses were kept anonymous. The study did not involve any sensitive personal data, and participants were free to exit the survey at any stage.

Result:

A total of 27 valid responses were collected through the online survey. Participants represented different age groups and levels of familiarity with artificial intelligence, ranging from slightly familiar to highly familiar. Most respondents reported having prior experience using AI-based tools such as chatbots, recommendation systems, or virtual assistants.

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(c) Jan – Feb, 2026

OPEN ACCESS Original Research Article

A. Comparison of Trust and Perception Scores

Participants rated two AI systems—System A (high accuracy without explanation) and System B (explainable AI with slightly lower accuracy)—across four dimensions: trust, perceived fairness, confidence, and willingness to rely on the system. Responses were recorded on a five-point Linear scale.

The results show that AI System B received higher average scores across all measured dimensions when compared to AI System A. As shown in Figure 1, the mean trust score for System A was 2.93, while System B achieved a higher mean trust score of 3.26. Figure 2 illustrates that perceived fairness increased from 3.11 for System A to 3.48 for System B.

Confidence in the system's decisions was also higher for the explainable system, with System A receiving a mean score of 3.15, compared to 3.37 for System B. Willingness to rely on the system followed a similar pattern, with System A scoring 3.15 and System B scoring 3.19. These findings suggest that the inclusion of explanations positively influenced participants' perceptions, even when the system was described as slightly less accurate.

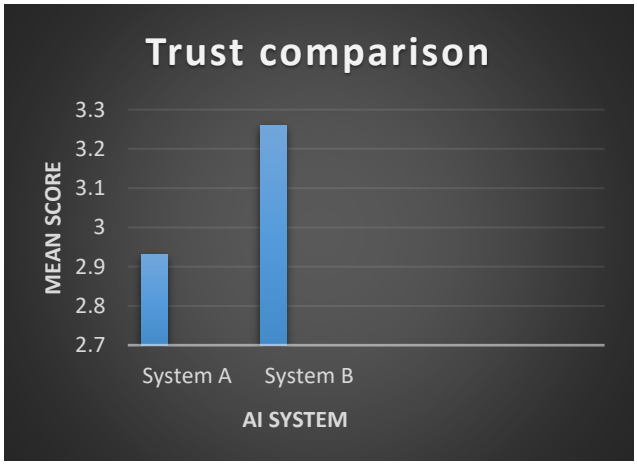

Figure 1. Comparison of mean trust scores between AI system A and AI System B

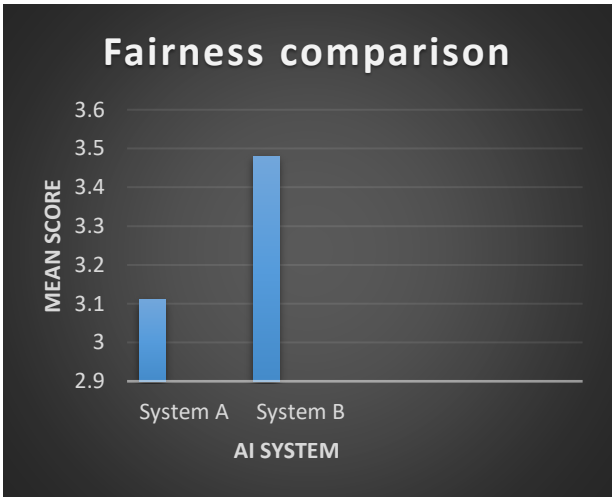

Figure 2. Comparison of perceived fairness scores for AI System A and AI System B

B. Preference for AI Systems in Real-Life Use

When asked which system they would prefer to use in real-life decision-making scenarios, a clear majority of participants favored the explainable system. Figure 3 shows that twenty-one out of 27 respondents indicated a preference for AI System B, while only six participants preferred the high-accuracy but non-explainable System A.

This preference highlights the importance participants placed on transparency and understanding when interacting with AI systems, particularly in decision-oriented contexts.

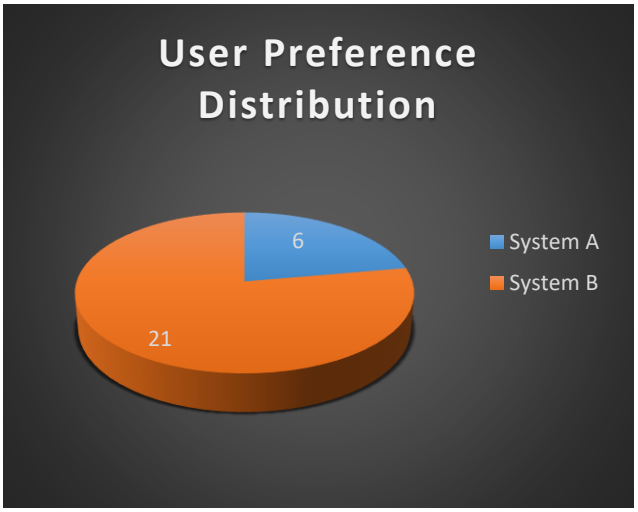

Figure 3. Participant preference between explainable and non-explainable AI systems

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(c) Jan – Feb, 2026

OPEN ACCESS Original Research Article

C. *Importance of Explainability and Qualitative Responses*

Responses to questions measuring the importance of explainability revealed that most participants rated it as an important or very important factor when evaluating AI systems. Open-ended responses further reinforced these findings. Many participants stated that explanations helped them feel more confident and reassured about the system's decisions, while some expressed discomfort with trusting a system that offered no reasoning, regardless of its accuracy.

Several participants also noted that explainability made the AI system feel more fair and user-friendly, increasing their willingness to accept and rely on such systems in the future.

Discussion:

The findings of this study highlight the significant role of explainability in shaping users' trust and acceptance of AI systems. Although AI System A was described as more accurate, participants consistently rated AI System B higher across trust, perceived fairness, and confidence. This suggests that users do not evaluate AI systems solely based on performance metrics, but also on how transparent and understandable the system appears to be.

One possible explanation for this preference is that explanations reduce uncertainty. When users are provided with reasons behind an AI system's decision, they are more likely to feel informed and in control, which enhances trust. In contrast, black-box systems may generate discomfort or skepticism, even when their accuracy is high, as users are unable to understand or question the decision-making process.

The strong preference for the explainable system observed in the results aligns with prior research emphasizing the importance of transparency and interpretability in human–AI interaction. Studies in explainable AI suggest that users are more willing to rely on AI systems when they can comprehend how decisions are made, particularly in contexts involving judgment or evaluation. The current findings reinforce these perspectives by demonstrating that explainability can outweigh marginal differences in accuracy from the user's point of view.

Additionally, qualitative responses revealed that participants associated explanations with fairness and accountability. This indicates that explainable systems may be perceived as more ethical and user-centered. Such perceptions are especially relevant in domains where AI decisions have direct consequences for individuals, such as education, healthcare, or finance.

Overall, the discussion suggests that designing AI systems with a focus on explainability can improve user trust and acceptance. While accuracy remains important, these findings indicate that transparency plays a crucial role in determining whether users are willing to rely on AI-driven decisions in practice.

Limitations

While this study provides useful insights into the role of explainability in shaping trust in AI systems, several limitations should be acknowledged.

First, the sample size was relatively small, with responses collected from only 27 participants. Although sufficient for an exploratory study, this limits the generalizability of the findings. A larger and more diverse participant pool would allow for stronger conclusions and improved statistical reliability.

Second, the study relied primarily on descriptive statistical analysis. While mean comparisons highlighted trends in user preference and trust, inferential statistical tests were not applied. As a result, the observed differences between the explainable and non-explainable AI systems cannot be claimed as statistically significant.

Third, the AI systems used in the experiment were simulated rather than real-world deployed models. Participants evaluated hypothetical descriptions of AI

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(c) Jan – Feb, 2026

OPEN ACCESS Original Research Article

behavior rather than interacting with actual systems. This may influence how trust and confidence are formed compared to real-life usage scenarios.

Fourth, the difference in accuracy between the two systems was described qualitatively as "slightly lower" rather than being quantified with specific numerical values. Participants may have interpreted this difference subjectively, potentially affecting their evaluations.

Finally, the study focused on a single decision-making context within the educational domain. Trust perceptions may vary across different application areas such as healthcare, finance, or law enforcement. Therefore, the findings may not fully generalize to all AI-driven decision-making scenarios.

Despite these limitations, the study offers valuable preliminary evidence that explainability plays a critical role in human trust and acceptance of AI systems, and it provides a foundation for future research in this area.

Future Work:

Future research can extend this study in several meaningful ways. First, increasing the sample size and recruiting participants from more diverse demographic and professional backgrounds would improve the generalizability of the findings. A larger dataset would also enable the use of inferential statistical methods to rigorously test the significance of differences in trust and perception between explainable and non-explainable AI systems.

Second, future studies could involve interactive, real-world AI systems rather than simulated descriptions. Allowing participants to directly engage with AI models and observe their behavior over time may provide deeper insights into how trust evolves through repeated use and real decision outcomes.

Third, future research could examine the trade-off between accuracy and explainability using explicit quantitative values. Presenting participants with precise accuracy levels would reduce ambiguity and help clarify how much accuracy users are willing to sacrifice in exchange for greater transparency.

Finally, extending the study across multiple application domains such as healthcare, finance, hiring, or criminal justice would provide a more comprehensive understanding of how contextual factors influence trust in AI systems. Such investigations would help inform domain-specific guidelines for designing trustworthy and human-centered AI systems.

Conclusion:

This study examined how explainability influences user trust and preference in AI-driven decision-making systems. Through a comparative survey-based experiment, the findings demonstrated that participants consistently favored an explainable AI system over a higher-accuracy black-box (AI)system. The results indicate that transparency and interpretability play a crucial role in shaping users' perceptions of trust, fairness, and confidence.

Even when informed that an AI system may be slightly less accurate, participants showed a clear preference for systems that provided understandable explanations for their decisions. This suggests that users value the ability to comprehend and evaluate AI outputs, rather than relying solely on performance metrics. The qualitative feedback further reinforced the importance of explainability in enhancing perceived fairness and accountability.

Overall, the findings highlight that designing AI systems with explainability in mind can significantly improve user acceptance and trust. While accuracy remains an essential component of AI performance, this study suggests that explainability is a key factor in determining whether users are willing to rely on AI systems in real-world scenarios. Future research could extend this work by exploring different application domains and larger, more diverse participant groups to further examine the balance between accuracy and explainability in AI systems.

ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(c) Jan – Feb, 2026

OPEN ACCESS Original Research Article

References:

1. *F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.*
2. *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, San Francisco, CA, USA, 2016, pp. 1135–1144.*
3. *Z. C. Lipton, "The mythos of model interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, Oct. 2018.*
4. *T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1–38, Feb. 2019.*
5. *A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges," Information Fusion, vol. 58, pp. 82–115, Jun. 2020.*
6. *R. Guidotti et al., "A survey of methods for explaining black box models," ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, Jan. 2019.*
7. *D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," AI Magazine, vol. 40, no. 2, pp. 44–58, Summer 2019.*

Cite This Article:

Nadar K.K.(2026). *Explainability Over Accuracy: A Human-Centered Study of Trust in Artificial Intelligence*. **In Aarhat Multidisciplinary International Education Research Journal**: Vol. XV (Number I, pp. 116–122)

Doi: https://doi.org/10.5281/zenodo.18641697