Global Journal of Statistics
Vol. 1 No.1 2026 Page 30-39
E-ISSN: XXXX-XXXX, DOI: 10.XXXXX/XXXX

The Impact of Model Interpretability on Decision Validity in Machine Learning Applications

Sinta Firmansyah[1]*, Andi Setiawan[2], Mega Hakim[3]
[1,2] Department of Mathematics Education, Faculty of Teacher Training and Education, Universitas Pendidikan Indonesia, Bandung, Indonesia
[3] Department of Educational Technology, Faculty of Education, Universitas Negeri Malang, Malang, Indonesia

Article Info

Received : 7 February 2026
Revised : 31 March 2026
Accepted : 16 April 2026

Keyword:

Model interpretability;
Decision validity;
Machine learning;
Explainable artificial intelligence;
Human-centered ai.

ABSTRACT

This study aims to explore the impact of model interpretability on decision validity in machine learning applications. The study responds to the growing use of machine learning in decision-support systems and the concern that accurate models do not always produce valid decisions when users cannot understand or evaluate their outputs. A qualitative instrumental case study approach was used to examine how users interpret model explanations, assess recommendation reliability, and justify machine learning-assisted decisions. Data were collected through semi-structured interviews, limited participant observation, and document analysis involving users, analysts, decision-makers, and technical staff who had direct experience with machine learning applications. Thematic analysis identified four main themes: interpretability as a basis for trust calibration, explanation gaps in decision validation, human judgment as a corrective mechanism, and organizational context as a condition for valid decisions. The findings show that interpretability supports decision validity when users can connect model explanations with contextual knowledge, professional judgment, and accountability procedures. However, interpretability does not automatically improve decision quality when explanations are too technical, unclear, or disconnected from user tasks. This study contributes to human-centered explainable artificial intelligence by framing interpretability as a practical, cognitive, and organizational process. The findings imply that organizations should develop clear review procedures, user-centered explanation designs, and governance mechanisms for responsible machine learning use. Future research should compare decision contexts across sectors and examine how different explanation formats affect user judgment.

***) Corresponding Author:**

Sinta Firmansyah,
Department of Mathematics Education, Faculty of Teacher Training and Education, Universitas Pendidikan Indonesia, Bandung, Indonesia
Email: sinta.firmansyah059@gmail.com

1. Introduction

Machine learning has become a major decision-support mechanism in health care, education, finance, public administration, business analytics, and digital governance. These systems process large datasets, detect hidden patterns, and generate recommendations that influence practical decisions. Yet high predictive accuracy does not always produce valid decisions when users cannot understand how a model reaches its output. Arrieta et al. (2020) argue that explainable artificial intelligence is needed because black-box systems create problems of transparency,

accountability, and trust. Ali et al. (2023) also show that explainability has become central to trustworthy AI because users need understandable reasons before they accept algorithmic recommendations. Linardatos et al. (2021) add that interpretability methods help users examine how machine learning models behave, but each method has different limits.

The problem becomes more serious when machine learning supports high-stakes decisions. In clinical settings, for example, a prediction can affect diagnosis, treatment priority, or resource allocation. Markus et al. (2021) explain that the demand for explainability depends on the user's purpose, such as verification, justification, learning, or accountability. Ghassemi et al. (2021) warn that explanation tools may create false confidence if users cannot assess whether the explanation reflects the real model behavior. Tjoa and Guan (2021) also show that explainable AI in medical applications still faces challenges related to reliability, usability, and clinical meaning. These findings indicate that model interpretability needs more than technical improvement. It must support human judgment in real decision contexts.

In practice, users often interact with machine learning outputs through dashboards, risk scores, automated alerts, ranking systems, or recommendation interfaces. These tools can shape user trust, but they can also create overreliance when users accept predictions without critical evaluation. Alufaisan et al. (2021) found that AI recommendations may improve decision accuracy in some contexts, but explanations do not always improve human decision quality. Poursabzi-Sangdeh et al. (2021) also found that transparent models do not automatically help users detect model errors. Buçinca et al. (2021) show that cognitive forcing strategies can reduce overreliance by encouraging users to think before accepting AI advice. These studies suggest that decision validity depends on how users interpret, question, and integrate model explanations into their own reasoning.

The issue of interpretability also has social and organizational relevance. Machine learning systems do not operate in neutral spaces. They shape professional responsibility, institutional accountability, and public trust. Langer et al. (2021) emphasize that explainable AI must address the needs of different stakeholders, including developers, decision-makers, affected individuals, and regulators. Ehsan et al. (2021) argue that AI explanation should include social transparency because users need to understand not only what the system predicts, but also how it fits into human work practices. Meske et al. (2022) state that XAI research needs clearer attention to objectives, stakeholders, and organizational implementation. Therefore, interpretability should be studied as a human-centered process, not only as a technical property of the model.

Recent studies show that human-centered evaluation remains an important gap in explainable AI research. Kim et al. (2024) found that many XAI studies still focus on system-level evaluation rather than users' lived experience with explanations. Rong et al. (2024) also show that user studies are needed to understand how explanations influence trust, understanding, satisfaction, and decision behavior. Leichtmann et al. (2023) found that explanations can improve trust calibration and decision behavior in high-risk tasks, but Leichtmann et al. (2024) show that the effect of explanation depends on task design, user background, and interaction context. These findings create a clear literature gap. Previous studies have examined explanation methods and decision performance, but fewer studies have explored how users construct meaning around decision validity when they work with interpretable machine learning systems.

Based on this gap, this study aims to explore the impact of model interpretability on decision validity in machine learning applications. The study focuses on users' experiences in understanding model explanations, assessing prediction reliability, identifying possible errors, and deciding whether a model-supported recommendation is valid in practice. The study contributes to human-centered explainable AI by examining interpretability as a practical, social, and cognitive process. It also offers practical value for system designers, data analysts, managers, and policy makers who need machine learning systems that support accountable decisions. Kaur et al. (2020) show that interpretability tools must match the real work practices of users, while

Liao et al. (2020) stress that explanation design should respond to users' actual questions. Van Leersum and Maathuis (2025) further argue that human-centered explainable decision-making requires attention to context, values, and user involvement.

2. Materials and Methods

In the method section, the author(s) should provide a detailed account of the procedures, materials, and equipment used in the research. This includes a comprehensive description of the study design, participants or subjects, materials used, procedures followed, and data analysis techniques employed. The goal is to offer sufficient information for other researchers to replicate the study, thereby ensuring the reproducibility and validity of the findings.

This study used a quantitative explanatory research design to test the effect of model interpretability on decision validity in machine learning applications. An explanatory design was selected because the study aimed to measure the relationship between variables and test whether model interpretability significantly predicts decision validity. The independent variable was model interpretability, while the dependent variable was decision validity. The study also considered user experience with machine learning as a control variable because prior research shows that user knowledge can influence how explanations are interpreted. Quantitative design was suitable because previous XAI studies have measured interpretability, trust, reliance, and decision outcomes through structured experimental and survey instruments (Alufaisan et al., 2021). This approach also aligns with human-AI decision-making studies that examine how explanation design affects user judgment through measurable indicators (Wang & Yin, 2022). In addition, the design reflects the need for systematic and reproducible XAI evaluation methods (Nauta et al., 2023). The use of an explanatory model was relevant because interpretability is expected to influence users' ability to make valid decisions from AI recommendations (Schemmer et al., 2023).

The study collected primary data through a structured questionnaire and a machine learning-assisted decision simulation. The questionnaire measured respondents' perceptions of explanation clarity, feature transparency, logical coherence, explanation usefulness, decision confidence, decision consistency, and perceived decision validity. The simulation presented respondents with several AI-assisted decision scenarios, such as risk classification, recommendation assessment, or prediction-based choice tasks. Respondents evaluated model recommendations with short explanation outputs, then selected decisions based on the information provided. This procedure followed prior studies that used AI-assisted decision tasks to examine reliance, overreliance, and explanation effectiveness (Buçinca et al., 2021). It also reflected experimental work showing that explanation engagement can influence whether users accept or reject AI advice (Vasconcelos et al., 2023). The use of explanation-based decision scenarios was relevant because explanation modality can affect decision accuracy and user reliance (Robbemond et al., 2022). The combination of questionnaire and simulation allowed the study to capture both perception-based and task-based decision validity (Leichtmann et al., 2023).

The population of this study consisted of users who had experience using, studying, or evaluating machine learning-based systems in academic, business, education, finance, health care, information systems, or data analytics contexts. The target respondents included students in technology-related programs, lecturers, data analysts, business professionals, and decision-makers who had basic knowledge of AI-assisted decision systems. The sample was selected using purposive sampling because respondents needed to meet specific criteria. These criteria included familiarity with machine learning applications, ability to read short AI explanations, and willingness to complete the questionnaire and simulation. Purposive sampling was appropriate because interpretability studies require respondents who can meaningfully evaluate AI recommendations and explanations (Kaur et al., 2020). Prior XAI research also shows that explanation needs differ across users, roles, and contexts (Liao et al., 2020). Dhanorkar et al. (2021) emphasized that different stakeholders require different explanation types across the AI

lifecycle. Therefore, the selected respondents needed enough exposure to AI systems to provide valid responses about model interpretability and decision validity (Ehsan et al., 2021).

The minimum sample size was set at 150 respondents to support regression analysis and provide sufficient statistical power for testing the research hypothesis. A larger sample was preferred when possible because perception-based and task-based measurements require stable estimates across respondent groups. The study used an online survey format to reach respondents from different professional and academic backgrounds. Screening questions were used at the beginning of the questionnaire to ensure that only eligible respondents participated. This procedure was consistent with prior quantitative studies in human-AI decision-making that used online experiments and structured decision tasks with screened participants (Buçinca et al., 2021). Wang and Yin (2022) also used controlled decision tasks to compare how different explanation principles affect human decisions. Robbemond et al. (2022) used a between-subjects design to examine the role of explanation modality in AI-assisted decision-making. The sampling and screening procedures helped ensure that the collected data reflected respondents who could evaluate machine learning explanations in a meaningful way (Schoeffer et al., 2024).

The research instrument consisted of three sections. The first section collected demographic data, including age, education level, field of work or study, experience with AI tools, and frequency of using machine learning-based applications. The second section measured model interpretability through indicators of clarity, transparency, understandability, feature relevance, and explanation usefulness. The third section measured decision validity through indicators of decision accuracy, decision consistency, justification quality, confidence calibration, and appropriate reliance. The instrument used a five-point Likert scale ranging from 1, strongly disagree, to 5, strongly agree. The development of interpretability indicators was based on XAI literature that defines interpretability as the degree to which humans can understand the logic and reasons behind model outputs (Arrieta et al., 2020). The decision validity indicators were informed by research on appropriate reliance and human-AI decision quality (Schemmer et al., 2023). The explanation quality indicators followed the view that XAI evaluation should include clarity, correctness, usefulness, and user-centered interpretation (Nauta et al., 2023). The instrument also reflected the need to measure explanation effects beyond general trust by examining actual decision judgment (Alufaisan et al., 2021).

Instrument validity was tested through content validity and construct validity. Content validity was conducted by asking three experts in machine learning, information systems, and quantitative research methodology to assess the relevance, clarity, and theoretical alignment of each item. Items that were unclear, redundant, or inconsistent with the research variables were revised before data collection. Construct validity was tested using exploratory factor analysis or confirmatory factor analysis, depending on the final sample size and data structure. Factor loading values above 0.50 were considered acceptable, while items below the threshold were reviewed or removed. This validation process was important because interpretability is a multidimensional construct that must be measured through clear and theoretically grounded indicators (Linardatos et al., 2021). Minh et al. (2022) emphasized that XAI evaluation must consider both technical and human-centered dimensions. Kim et al. (2024) also argued that human-centered XAI evaluation requires instruments that reflect user understanding and decision context. Therefore, validity testing ensured that the instrument measured the intended constructs accurately and consistently (Nauta et al., 2023).

Reliability testing was conducted using Cronbach's alpha and composite reliability. A reliability value of 0.70 or higher was considered acceptable for internal consistency. Before hypothesis testing, the data were examined through descriptive statistics, missing value checks, outlier detection, and assumption testing. The assumption tests included normality, linearity, multicollinearity, and heteroscedasticity tests. Descriptive statistics were used to present respondent profiles and variable tendencies through frequency, percentage, mean, and standard deviation. This procedure followed the quantitative tradition in AI-assisted decision studies that

combine perception measures with statistical testing of decision outcomes (Wang & Yin, 2022). Reliability testing was also needed because users may interpret XAI indicators differently across tasks and domains (Kaur et al., 2020). Prior XAI evaluation research stresses that explanation quality should be measured with systematic and reproducible methods (Nauta et al., 2023). The reliability and assumption tests therefore supported the credibility of the statistical analysis and reduced measurement error (Ali et al., 2023).

The main data analysis used Pearson correlation and simple linear regression to test the relationship between model interpretability and decision validity. Pearson correlation measured the direction and strength of the relationship between the two variables. Simple linear regression tested whether model interpretability significantly predicted decision validity. If additional variables such as trust calibration, AI experience, or explanation usefulness were included, multiple linear regression could be used to assess their combined effects. The hypothesis was tested at a significance level of 0.05. The null hypothesis stated that model interpretability has no significant effect on decision validity. The alternative hypothesis stated that model interpretability has a significant positive effect on decision validity. This analytical model was consistent with research showing that explanations can influence reliance, performance, and human judgment in AI-assisted decision-making (Leichtmann et al., 2023). It also reflected evidence that explanation effects may vary depending on user engagement and task difficulty (Vasconcelos et al., 2023). The regression approach helped estimate how much variance in decision validity could be explained by interpretability (Schemmer et al., 2023). All statistical analyses were conducted using SPSS, R, or SmartPLS, depending on the final structure of the measurement model and the complexity of the hypothesis testing (Schoeffer et al., 2024).

3. Results and Discussions

The qualitative analysis produced four main themes: interpretability as a basis for trust calibration, explanation gaps in decision validation, human judgment as a corrective mechanism, and organizational context as a condition for valid machine learning decisions. These themes emerged from semi-structured interviews, limited observation of system use, and review of internal documents related to machine learning-based decision support. The findings show that users did not define decision validity only as the accuracy of model output. They understood validity as the extent to which model recommendations could be explained, checked, compared with contextual knowledge, and justified before action. This pattern indicates that interpretability functions as a bridge between algorithmic prediction and accountable human decision-making.

The first theme shows that interpretability helped users calibrate trust in machine learning outputs. Participants stated that they were more willing to use a recommendation when the system explained which variables influenced the result. For example, one participant explained, “I do not need the model to show every formula, but I need to know why it gives this risk level” (P3). Another participant said, “When the dashboard shows the most influential factors, I can decide whether the prediction makes sense in the actual case” (P7). Observation also showed that users often paused before accepting recommendations and compared feature explanations with their own professional experience. This pattern suggests that interpretability did not simply increase trust. It helped users decide when to trust, question, or reject model output.

The second theme relates to explanation gaps in decision validation. Participants reported that some explanation features were too technical, too general, or not directly connected to the decision problem. Several users stated that charts, probability scores, or feature rankings were useful only when they could be linked to practical reasoning. One participant noted, “The explanation says which factor is important, but it does not always

tell me what I should do next" (P2). Another participant added, "Sometimes the model gives a high score, but the explanation does not match what we see in the field" (P5). Document analysis also showed that system manuals explained technical features, but they rarely explained how users should evaluate model recommendations in uncertain cases. This finding shows that interpretability can support decision validity only when explanations are relevant to the user's task, role, and decision responsibility.

The third theme indicates that human judgment remained central in validating machine learning decisions. Participants did not treat the system as an automatic authority. They used model explanations as one source of evidence alongside experience, institutional rules, and contextual information. One participant stated, "The model helps us see patterns, but the final decision still needs human review because the data does not always show the whole situation" (P4). Another participant said, "If the explanation conflicts with what we know from the case history, we discuss it before making a decision" (P8). Observation confirmed that users often discussed model results in teams, especially when predictions involved risk, fairness, or resource allocation. This pattern shows that decision validity emerged through interaction between algorithmic evidence and human interpretation, not from the model alone.

The fourth theme concerns organizational context. Participants explained that interpretability became more useful when the organization had clear procedures for reviewing, documenting, and challenging model-based recommendations. In units with informal review practices, users relied heavily on individual judgment. In units with clearer decision protocols, users could compare model output with guidelines and document reasons for accepting or rejecting recommendations. One participant stated, "The explanation is useful when we have a procedure for checking it. Without that, everyone interprets it differently" (P1). Another participant said, "We need a record of why the model recommendation was followed or not followed, especially for sensitive decisions" (P6). These findings suggest that interpretability is not only a feature of the machine learning model. It also depends on organizational routines, documentation practices, and accountability structures.

Overall, the findings show that model interpretability affects decision validity through three interconnected processes. First, it helps users understand the basis of model recommendations. Second, it enables users to compare algorithmic output with contextual knowledge. Third, it supports accountability when decisions must be explained to others. However, interpretability does not automatically produce valid decisions. Its impact depends on the clarity of explanation, user expertise, decision context, and organizational readiness. The data suggest that valid machine learning-assisted decisions require a human-centered system in which users can question, contextualize, and justify the use of model recommendations.

Discussion

The findings support the view that explainable artificial intelligence must be understood as a human-centered process. Arrieta et al. (2020) explain that XAI is needed to address transparency, trust, and accountability, while Ali et al. (2023) emphasize that explainability contributes to trustworthy AI when users can understand and evaluate model behavior. The present findings strengthen this argument by showing that users did not accept interpretability as a technical display alone. They valued explanations that helped them judge whether a decision was reasonable in a specific context. This aligns

with Langer et al. (2021), who argue that XAI should respond to the needs of different stakeholders, not only the needs of model developers.

The finding that interpretability helps trust calibration is consistent with previous studies on AI-assisted decision-making. Leichtmann et al. (2023) found that explainable AI can improve trust and human behavior in high-risk decision tasks. Rong et al. (2024) also show that model explanations influence user understanding, trust, and decision behavior. However, the present study adds a more contextual perspective. Participants did not describe trust as blind acceptance. They treated trust as a careful judgment formed through comparison between model explanations, professional experience, and field conditions. This finding extends the literature by showing that interpretability supports decision validity most strongly when it encourages critical trust rather than automatic compliance.

The findings also confirm that explanations can fail when they do not match user needs. Poursabzi-Sangdeh et al. (2021) found that transparent models do not always help users identify model errors. Alufaisan et al. (2021) also show that explanations do not always improve human decision quality. The current study explains why this may happen. Participants struggled when explanations were too technical, too abstract, or disconnected from practical action. In such cases, interpretability became a formal feature rather than a meaningful decision-support tool. This suggests that future XAI design should move beyond asking whether a model is explainable. Researchers and developers need to ask whether the explanation helps users validate decisions in real work situations.

The role of human judgment in this study also supports the argument that machine learning should not replace professional reasoning in complex decisions. Ghassemi et al. (2021) warn that current approaches to explainable AI in health care may create false confidence if users cannot assess the actual reliability of explanations. Buçinca et al. (2021) similarly argue that cognitive forcing functions can reduce overreliance on AI by encouraging users to think critically. The present study shows that users often protected decision validity by discussing model outputs, checking contextual information, and documenting disagreements with system recommendations. This finding indicates that interpretability works best when it strengthens human reasoning, not when it removes human responsibility.

The organizational dimension of the findings contributes to the discussion on accountability in machine learning applications. Ehsan et al. (2021) argue that AI systems require social transparency because explanation is shaped by the social context in which the system operates. Meske et al. (2022) also state that explainable AI must consider objectives, stakeholders, and organizational implementation. The current findings support these arguments by showing that interpretability becomes more useful when organizations provide review procedures, documentation standards, and clear responsibility structures. Without these structures, users may interpret model explanations inconsistently. This can weaken decision validity even when the model provides technically sound explanations.

The findings also have theoretical implications for human-centered explainable AI. Kim et al. (2024) argue that XAI evaluation should focus on users and their interaction with explanations. Liao et al. (2020) show that users ask different explanation questions depending on their goals, tasks, and uncertainty. This study contributes to that perspective by showing that decision validity depends on how users translate explanations into

practical judgment. Interpretability is therefore not only a model property. It is also a situated interpretive process shaped by user knowledge, task complexity, institutional norms, and ethical responsibility.

Practically, the findings suggest that organizations should not implement machine learning systems without explanation governance. System designers need to provide explanations that are clear, task-relevant, and connected to possible actions. Organizations also need procedures that guide users in reviewing, accepting, rejecting, and documenting model recommendations. Training should not only teach users how to read dashboards. It should also teach them how to question model output, identify possible bias, and combine algorithmic evidence with professional judgment. These steps can help reduce overreliance and improve the validity of machine learning-assisted decisions.

This study also opens several directions for future research. Future studies can compare different sectors, such as health care, education, finance, and public administration, to examine how decision context shapes the meaning of interpretability. Further research can also combine qualitative interviews with usability testing or decision simulation to evaluate how users respond to different explanation formats. In addition, future studies can examine the experiences of people affected by machine learning decisions, not only the professionals who use the systems. Such work would deepen understanding of interpretability as a social, ethical, and practical component of valid decision-making in machine learning applications.

4. Conclusions

This study concludes that model interpretability plays an important role in strengthening decision validity in machine learning applications. The findings show that users do not assess decision validity only from predictive accuracy, but also from their ability to understand, question, compare, and justify model recommendations. Four main themes emerged from the analysis: interpretability as a basis for trust calibration, explanation gaps in decision validation, human judgment as a corrective mechanism, and organizational context as a condition for valid machine learning decisions. These findings contribute to the literature on human-centered explainable artificial intelligence by showing that interpretability is not merely a technical property of a model, but a practical and social process that shapes how users make sense of machine learning outputs in real decision contexts.

The study also offers practical and policy implications. Organizations that use machine learning for decision support need clear procedures for reviewing, documenting, and challenging model recommendations. System designers should develop explanation features that are understandable, task-relevant, and connected to user decision needs. From a policy perspective, institutions should promote accountability standards that require human review, explanation transparency, and documentation of model-assisted decisions. Future research can expand this study by comparing different sectors, such as health care, education, finance, and public administration, or by combining qualitative inquiry with usability testing to examine how different explanation formats influence decision validity.

References

Ahmed, S. K. (2024). The pillars of trustworthiness in qualitative research. *Journal of Medicine, Surgery, and Public Health, 2*, 100051. https://doi.org/10.1016/j.glmedi.2024.100051

Ahmed, S. K., Mohammed, R. A., Nashwan, A. J., Ibrahim, R. H., Abdalla, A. Q., Ameen, B. M. M., & Khdhir, R. M. (2025). Using thematic analysis in qualitative research. *Journal of*

Medicine, Surgery, and Public Health, 6, 100198. https://doi.org/10.1016/j.glmedi.2025.100198

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion, 99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(17), 15238–15246. https://doi.org/10.1609/aaai.v35i17.17706

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in reflexive thematic analysis? *Qualitative Research in Psychology, 18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 188. https://doi.org/10.1145/3449287

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3411764.3445188

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Hong, S., & Park, W. (2025). Developing user-centered system design guidelines for explainable AI: A systematic literature review. *Artificial Intelligence Review, 58*, Article 386. https://doi.org/10.1007/s10462-025-11363-y

Johnson, J. L., Adkins, D., & Chauvin, S. (2020). A review of the quality indicators of rigor in qualitative research. *American Journal of Pharmaceutical Education, 84*(1), 7120. https://doi.org/10.5688/ajpe7120

Jung, J., Kang, S., Choi, J., El-Kareh, R., Lee, H., & Kim, H. (2025). Evaluating the impact of explainable AI on clinicians' decision-making: A study on ICU length of stay prediction. *International Journal of Medical Informatics, 201*, 105943. https://doi.org/10.1016/j.ijmedinf.2025.105943

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Vaughan, J. W. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376219

Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. *Frontiers in Artificial Intelligence, 7*, 1456486. https://doi.org/10.3389/frai.2024.1456486

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*, 103473. https://doi.org/10.1016/j.artint.2021.103473

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior, 139*, 107539. https://doi.org/10.1016/j.chb.2022.107539

Leichtmann, B., Hinterreiter, A., Humer, C., Streit, M., & Mara, M. (2024). Explainable artificial intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival. *International Journal of Human–Computer Interaction, 40*(10), 2534–2549. https://doi.org/10.1080/10447318.2023.2221605

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3313831.3376590

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18. https://doi.org/10.3390/e23010018

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management, 39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. https://doi.org/10.1145/3411764.3445315

Rong, Y., Leemann, T., Nguyen, T. T., Fiedler, L., Qian, P., Unhelkar, V. V., Seidel, T., Kasneci, G., & Kasneci, E. (2024). Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(4), 2104–2122. https://doi.org/10.1109/TPAMI.2023.3331846

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Van Leersum, C. M., & Maathuis, C. (2025). Human centred explainable AI decision-making in healthcare. *Journal of Responsible Technology, 21*, 100108. https://doi.org/10.1016/j.jrt.2025.100108