Exploring Evaluation Methodologies for Explainable AI: Guidelines for Objective and Subjective Assessment

Sule Tekkesinoglu
Department of Engineering Science
University of Oxford
Oxford, UK
sule@robots.ox.ac.uk

Abstract

This article explores the landscape of evaluation methodologies for Explainable Artificial Intelligence (XAI), focusing on both objective and subjective assessment paradigms. Objective evaluation methods provide structured and measurable assessments, offering insights into the system’s performance, technical robustness, and functional effectiveness through quantitative metrics. These metrics are practical in benchmarking, verifying, and estimating functional effectiveness against predefined standards and existing explanation methods. Subjective assessments explore user-centric dimensions, elucidating perceptions, preferences, and interactions with AI-generated explanations. These assessment techniques unravel the human factors, capturing the user’s experience of quality, satisfaction, trust, and performance. This work presents a list of evaluation metrics concerning objective and subjective properties of explanations. Additionally, concerning subjective measurements, an example of a user satisfaction questionnaire along with open-ended questions is provided. These can be tailored based on the context of the explanations and the user group being surveyed. Taken together, the outlined assessment methods provide a framework that caters to the diverse facets of XAI explainability to facilitate a deeper understanding of XAI systems’ capabilities and enhance trust in AI-driven technologies.

1 Introduction

In the rapidly advancing landscape of Artificial Intelligence (AI), the quest for understanding and trust in AI decision-making processes has gained paramount importance. Explainable AI (XAI) has emerged as a critical area, aiming to demystify complex AI models and bridge the gap between machine reasoning and human comprehension. However, assessing the effectiveness of explainability methods remains a multifaceted challenge, encompassing objective and subjective dimensions (Silva et al. 2023; Lopes et al. 2022). These dimensions complement each other, offering a holistic assessment of XAI systems. This article delves into the diverse landscape of evaluation methodologies for explainable AI, studying measurements for objective metrics and subjective assessments (See Figure 1). By dissecting the duality between quantifiable performance indicators and nuanced user experiences, this exploration aims to untangle the intricacies of evaluating XAI methods.

Objective evaluation methods, rooted in quantitative metrics, provide structured and measurable assessments, offering insights into XAI system’s performance (Zhou et al. 2021; Nauta et al. 2023). These measures excel in benchmarking performance against predefined standards and verifying technical robustness (Le et al. 2023). They thrive in scenarios requiring empirical validation, offering precise and quantitative insights into XAI methods. Within this study, objective evaluation focuses on functional and safety evaluation metrics. Functional evaluation methods gauge the effectiveness of explanations based on their utility in addressing specific functions. Generally, there is consensus in the literature on performance metrics such as fidelity, consistency, and stability.
On the other hand, safety measures for explanations in AI systems are centred on guaranteeing accuracy, fairness, and robustness, especially in high-stakes scenarios. Current research underscores the need to incorporate safety and security measures for explanations into existing assessment practices.

Subjective assessments elucidate user perceptions, preferences, and interactions, providing contextual depth to evaluations. These methods excel in scenarios demanding user-centric design and understanding of human-AI interactions (Maxwell and Dumas [2023]; Chromik and Butz [2021]). Subjective evaluation methods delve into the nuanced realm of user experiences, capturing perceptions of quality, satisfaction, trust, and performance. Quality assessments encompass comprehensibility, justifiability, and interactivity of explanations, ensuring alignment with user expectations. User satisfaction metrics gauge user contentment, evaluating ease of use and usefulness of AI explanations. Given an explanation, trust evaluations determine the user’s confidence in the system’s decision-making process and recommendations. Performance metrics analyze the effectiveness and efficiency of users in completing tasks facilitated by explanations. Collectively—with the objective evaluations, these measurements offer a comprehensive perspective on the effectiveness, user acceptance, and robustness of AI-generated explanations.

Figure 1: Evaluations of explanation methods are categorized into objective and subjective criteria, each with relevant properties to be assessed.

The rest of the paper is organized as follows: Section 2 discusses the objective assessment under functional and safety measurement. Section 3 presents the subjective evaluation metrics by exploring quality, satisfaction, trust, and performance measurements. Finally, Section 4 concludes the paper.

2 Objective Evaluation

Objective evaluation includes objective metrics and automated approaches to assess methods for explainability. This type of evaluation uses certain properties of explainability as a proxy for explanation quality. The use of explainability properties to objectively evaluate the quality of
explanations has proven beneficial. In fact, these experiments can be particularly advantageous compared to human-subject studies, given that the need for funding and approvals to conduct such studies often exceeds the resources of an ML researcher. Objective evaluations are especially useful when one already has a validated class of baseline models to work with. They may also be relevant when a method is in its early stage of development or when human subject studies are inadmissible for ethical reasons (Doshi-Velez and Kim [2017]). In such cases, a combination of theoretical analysis, empirical testing, and specialized tools can provide a robust understanding of the quality of generated explanations from an objective standpoint (Vilone and Longo [2020]; Robnik-Šikonja and Bohanec [2018]). This section examines objective evaluation through the lenses of functional and safety and security evaluation metrics.

2.1 Functional Evaluation Metrics

Functional evaluation methods assess the effectiveness of explanations based on their functional utility. These methods evaluate explanations through quantitative measures using mathematical specifications and tests in empirical studies without relying on human input. The evaluation criteria consider factors such as faithfulness to the model, consistency across multiple models, and stability to minor changes in input. These factors are discussed in further detail below.

2.1.1 Fidelity

This is associated with how well the interpretation method agrees with the input-output mapping of the model and reflects the global relationships learned by the model. An explanation with low fidelity means it does not approximate well to the original model and cannot furnish valid reasons given that the input-output mapping is incorrect. High fidelity is always desirable regardless of the model’s accuracy. If the model has high accuracy and the explanation has high fidelity, the explanation hence has high accuracy. However, low explanation accuracy is expected if the accuracy of the machine

Figure 2: Objective evaluation through functional and safety and security evaluation metrics
learning model is likewise low (Carvalho et al. [2019]; Ras et al. [2018]). There are several ways to assess explanation fidelity, which include:

**Model Understanding** – This involves comparing the explanation’s content to the model’s internal functioning and checking if the explanation aligns with the actual decision-making process of the AI model. This process helps to measure its reliability and trustworthiness.

**Local and Global Coherence** – This concerns whether local explanations align consistently with the model’s overall behaviour across various data samples. It also involves studying the interdependencies of features, as a globally important feature might not have the same level of importance in individual predictions.

**Adversarial Testing** – It refers to testing the robustness of explanations by introducing adversarial samples or edge cases. This helps evaluate whether the explanation adapts sensibly to these cases and accurately reflects the model’s reasoning.

### 2.1.2 Consistency

Consistency relates to how an explanation differs between models trained on the same task with similar output predictions. If the explanations indicate the same features in a similar degree of importance, then the explanations are highly consistent. It is important to note that models that differ significantly in their architecture and learning mechanisms are less likely to produce consistent explanations for the same instances (Carvalho et al. [2019]). It is described as the ‘Rashomon Effect’ in which an event is given contradictory interpretations or descriptions by the observers (i.e., models) involved (Breiman [2001]). Therefore, the degree of relevance for high consistency in explanations across models depends upon the similarity in how these models process and interpret data. Models with shared or similar data-driven relationships are expected to produce consistent explanations, while models with disparate underlying mechanisms may yield divergent explanations. Here are some factors to consider:

**Scope of Comparison** – Defining the scope of comparison, whether across different models or variations of the same model, is crucial to accurately measuring the consistency of explanations.

**Instances for Comparison** – It is necessary to select instances or cases for comparison that cover a diverse range of scenarios or complexities representative of the dataset. This can also entail cases where the model has made errors.

**Similarity Metrics** – Establishing appropriate evaluation metrics is essential to quantify the similarity between explanations and determine whether observed consistency is statistically significant or occurs by chance. Depending on the nature of the explanations, metrics such as cosine similarity, overlap measures, or specialized similarity indices may be appropriate.

### 2.1.3 Stability

Stability indicates the consistency or robustness of an explanation method concerning variations in input data. Consistency compares explanations between different models, whereas stability compares explanations between similar instances for a specified model. It signifies high stability when explanations remain unchanged or exhibit minimal variation with slight alterations in input features within similar instances. High stability is always desirable; however, a lack of stability may originate from non-deterministic elements of the explanation method (e.g., data sampling and perturbation step) (Carvalho et al. [2019]). Several standard methods exist to measure stability, including:

**Perturbation Analysis** – Introducing slight changes or perturbations to the input data and observing the changes in the explanations.

**Jaccard or Similarity Index** – Using similarity measures like the Jaccard index or cosine similarity to compare the similarity between explanations for slightly modified instances.

**Statistical Tests** – Applying statistical tests (e.g., t-tests, ANOVA (Analysis of Variance), Chi-square tests) to assess the significance of differences in explanations between similar instances.
2.1.4 Generalizability

Generalizability describes the range of models to which the explanation method can be applied. Model-agnostic methods are the highest in generalizability. The broad applicability of an explanation method increases the practicality, allowing for the possibility of assessing the consistency of explanations in a diverse group of models (Anjomshoae et al. [2019]; Ribeiro et al. [2016]). However, model-agnostic approaches might lack the specificity or nuanced insight into certain model intricacies that model-specific explanation methods can provide. Balancing between generalizability and specificity becomes crucial when choosing an explanation method based on the context and the need for insights into model-specific behaviour versus broader model applicability. Here are a few approaches to evaluate generalizability:

**Cross-Model Consistency** – To evaluate the generalizability, one can apply the explanation method to different models with varied architectures and assess the consistency of the explanations for similar instances. Higher consistency signifies better generalizability.

**Performance Metrics** – Quantifying the method’s performance, such as accuracy, stability, or fidelity, across multiple models provides insight into generalizability. A method showing consistent high performance across various models suggests good generalizability.

2.1.5 Explanatory Power

Explanatory power relates to the range of questions that the method can address, such as why, why-not, and what-if scenarios, as well as the various ways in which it can provide answers, including visuals, verbal responses, and decision trees. Moreover, the explanation method should consider the user’s needs and specifications, as emphasized in Ras et al. [2018]. Explanatory power also links to expressiveness in the number of events that can be explained and the ways that can generate explanations, i.e., language or structured explanations. It also relates to the idea that the explainer should be able to take both local and global perspectives, preferably explaining individual predictions along with the model’s overall behaviour (Ribeiro et al. [2016]). Measuring explanatory power involves evaluating the breadth and depth of an explanation method’s capabilities, which include:

**Event Coverage** – Evaluating the method’s capacity to explain different events or instances, encompassing both specific cases and general model behaviour.

**Comparative Analysis** – Comparing the method’s performance against other explanation methods in terms of its range and depth of explanations provided.

2.1.6 Algorithmic Complexity

Algorithmic complexity refers to the level of computational difficulty or the resources required in computing or generating explanations (Carvalho et al. [2019]). The time taken to generate explanations becomes a limiting factor due to the algorithm’s complexity. Assessing algorithmic complexity involves evaluating the computational resources an algorithm consumes relative to its input size. There are several metrics that help assess algorithmic complexity, which include:

**Time Complexity** – Analyzing how the computation time scales with the input size.

**Space Complexity** – Evaluating the amount of memory or space the algorithm requires concerning its input.

**Empirical Testing** – Measuring the actual execution time or memory consumption of the algorithm with different input sizes.

**Benchmarking** – Comparing the algorithm’s performance against known benchmarks or standard algorithms with established complexities.

2.2 Safety and Security Evaluation Metrics

Safety and security measures for explanations in AI systems primarily involve ensuring that the provided explanations are accurate, fair, and robust, particularly in critical applications such as autonomous vehicles (Sokol and Flach [2020]). Evaluating information leakage concerning predictive models and training data and examining the consistency of explanations across various models are...
key aspects. While safety evaluation in explainability aims to assess robustness and security, precise quantitative metrics for these aspects in explanations are still evolving in the literature. This absence underscores the ongoing need for exploration and standardization in safety and security evaluation metrics for explainability methods. Here are potential metrics to consider for safety evaluation:

### 2.2.1 Information Leakage

Information leakage refers to revealing sensitive or excessive information by explaining the model or its training data. Metrics such as entropy and information gain/loss functions can help measure information leakage. Several techniques can be used to quantify information leakage, including:

**Entity Recognition** – This helps identify and flag sensitive entities like personal names, locations, financial details, or confidential information. Recognizing these entities helps assess the extent to which such sensitive information is disclosed in the explanation.

**Privacy Metrics** – Employing privacy-focused measures like differential privacy mechanisms to quantify the amount of sensitive information disclosed.

### 2.2.2 Fairness and Bias Metrics

Fairness and bias evaluation ensures that the provided reasoning or justifications are unbiased and impartial across different demographics, characteristics, or groups. Bias in explanations can occur if the AI system’s reasoning is influenced by certain characteristics like race, gender, or other sensitive attributes, leading to unfair or discriminatory explanations. Some of the key steps and considerations that must be taken into account during the evaluation process include:

**Root Cause Analysis** – Understanding the causes of biases in explanations by exploring how biases are introduced, whether through training data, algorithms, or systemic issues in the AI system.

**Bias Detection** – Employing techniques to identify biases present in the explanations by analyzing whether certain groups are consistently favoured or disadvantaged in the explanations provided by the AI system.

**Evaluation Metrics** – Defining metrics to assess fairness in explanations such as statistical parity, disparate impact analysis, or equalized odds to gauge if explanations are balanced across different demographic groups.

**Misrepresentation Detection** – Using techniques to identify if explanations misrepresent the AI model’s decision-making process, either unintentionally due to limitations in the explanation method or deliberately when the explanation is manipulated.

### 3 Subjective Evaluation

Human-grounded evaluation, also known as human-centred evaluation (Doshi-Velez and Kim [2017]), is a human-in-the-loop evaluation approach that utilizes end-users’ feedback and their informed opinion (Vilone and Longo [2020]; Hoffman et al. [2023]). Human-grounded experiments encompass two groups: laypersons (i.e., novice users) and domain experts. Studies involving laypeople are easily accessible since they require no prior technical/domain knowledge, allowing for a larger subject pool. On the other hand, conducting experiments with highly trained domain experts is more challenging due to the difficulty of accessing and compensating for their expertise. Nonetheless, human-centred studies may require domain experts when their informed judgment on the explanations produced by the model is necessary to verify the consistency of the explanations with the domain knowledge.

Human-centred studies involve users interacting with AI systems to assess the quality of explanations, user satisfaction, and trust (Speith and Langer [2023]). This process includes gathering both qualitative and quantitative feedback through surveys, interviews, or observational studies. The assessment of XAI models with human subjects can be categorized in the following ways.

#### 3.1 Quality of Explanations

Looking across the literature on explanations, there is a consensus on what makes for a good quality explanation from the perspective of social sciences (e.g., comprehensibility, selection, and social
Subjective evaluation methods explore user-centric dimensions such as comprehensibility of explanations, satisfaction, trust, and performance (Confalonieri et al. [2021]; Miller [2019]). As a result, it becomes possible to assess the quality of a given explanation based on these established criteria. The subjects who evaluate a particular AI explanation should be someone other than the ones who created the system in a valid experiment. This allows for an independent and unbiased assessment of the explanations generated by the system. Following sections delve into the various sub-properties that contribute to a good-quality explanation in more detail below.

### 3.1.1 Comprehensibility

Comprehensibility is related to how well humans understand the explanations. However, comprehension is subjective and dependent on the audience and context (Hoffman et al. [2018]). Compre-
hensibility is particularly relevant in safety-critical applications where ambiguity must be avoided (Ras et al. [2018]). Clear and concise explanations that explicitly address critical causes while avoiding information overload are key to achieving comprehensibility (Vilone and Longo [2020]). Tests and analyses commonly conducted in human subject studies to assess the comprehensibility of explanations in AI systems are as follows:

Cognitive Load Analysis – Assessing the cognitive effort and the time users are required to understand and utilize explanations provided by AI systems. This evaluation helps determine if explanations are clear and easily processed by users.

Eye-Tracking Studies – Observing and analyzing where users focus their attention within an explanation interface to understand which parts they find most relevant or confusing.

A/B Testing – Comparing different versions of explanations to determine which format, style, or content is more effective in conveying information and aiding user understanding. It also includes testing users’ understanding of explanations with varying lengths to see if shorter explanations are equally informative.

Error Analysis – Examining instances where users misinterpret or misunderstand explanations to identify common errors or areas of confusion.

3.1.2 Justifiability

Justifiability pertains to the extent to which an expert can evaluate the explanations provided by an AI model to ascertain if the model’s decisions align with established domain knowledge or expertise in the field. It requires domain experts to evaluate the explanations and confirm that the model adheres to known principles, rules, or expected behaviour within that domain. This validation by experts helps ensure that the AI model’s reasoning is consistent with the established knowledge or expertise in the relevant field (Biran and Cotton [2017]). Evaluating justifiability with experts involves several steps, including:

Domain Expert Selection – Identifying experts in the field or domain relevant to the AI model’s application. These experts should have a deep understanding of the subject matter.

Assessment Criteria – Providing guidelines or criteria for the experts to evaluate the explanations. This might include assessing the alignment of the AI model’s decisions with established domain principles, industry standards, or expected behaviour.

Verification Process – Allowing the experts to verify the explanations provided by the AI model. They might compare the AI-generated explanations with known domain knowledge or verify if the AI’s decisions align with established practices.

Feedback Collection – Gathering feedback from the experts regarding the justifiability of the explanations. This feedback might include their assessment of the model’s alignment with domain knowledge, potential discrepancies, or areas needing improvement.

Validation and Consensus – Validating the expert assessments through discussions or consensus-building sessions among multiple experts. This step helps ensure a collective and validated judgment on the justifiability of the AI model’s explanations.

3.1.3 Interactivity

Interactivity refers to the capacity of an explanation method to reason about prior interactions to interpret and respond to users’ follow-up questions (Madumal et al. [1903]). In essence, an interactive explanation method should be capable of adapting and responding contextually to user inputs and inquiries, enhancing the overall user experience. Considering the following aspects helps in a comprehensive evaluation of the interactivity level in an explanation system:

Response Time – How quickly does the system provide follow-up explanations or respond to user queries?

Adaptability – Can the system adapt its explanations based on the user’s previous interactions or changing contexts?

User-Friendliness – How intuitive and user-friendly are the system’s interactions and responses?

Electronic copy available at: https://ssrn.com/abstract=4667052
Error Handling – How does the system manage errors or handle situations where it cannot provide a satisfactory explanation?

Feedback Mechanisms – Are there mechanisms for users to provide feedback, and are they effectively incorporated to improve interactivity?

Scalability – Can the system maintain its level of interactivity as the complexity or volume of interactions increases?

3.2 Explanation Satisfaction

Explanation satisfaction is the degree to which users feel the ease of use and usefulness of an AI system or process. Even though an explanation might be considered ‘good quality’ in the way described above, it may at the same time not be adequate or satisfying to users in a given context. The critical attributes of satisfactory explanations include but are not limited to:

Completeness – This relates to the ability of an explanation method to describe the underlying system sufficiently (Kulesza et al. [2013]). Oversimplification of statements may be detrimental to users’ trust. This measure determines if the explanation covers all relevant aspects adequately.

Actionability/Persuasiveness – It is related to the capacity of an explanation method to transfer convincing knowledge to end-users that the system’s decisions are actionable. An actionable explanation not only provides information but also prompts the user toward a specific course of action or decision (Kulesza et al. [2013]).

Usefulness – This refers to the extent to which an explanation assists users in decision-making or understanding. The usefulness of explanations is closely linked to the balance between their effectiveness and efficiency. An explanation might be highly effective but inefficient, offering comprehensive insights yet overly complex or time-consuming. Conversely, an explanation might be very efficient but ineffective, being too simplistic or failing to adequately address user needs (Tintarev and Masthoff [2015]).

Relevance – The relevance of explanations delves into how aptly the provided information addresses the specific question or situation the user raises (Nauta et al. [2023]). It assesses the direct applicability and alignment of the explanation with the user’s immediate needs or context, ensuring it is tailored and pertinent to the query or circumstances at hand.

Personalization – Personalization in explanations evaluates whether the provided information is customized or adapted to suit individual users’ specific needs, preferences, or characteristics (Schneider and Handali [2019]). It revolves around how the explanation is crafted or adjusted based on the user’s context, prior interactions, or known preferences, aiming to enhance relevance and engagement.

Timeliness – This relates to assessing whether the information is delivered at an appropriate moment or within a suitable timeframe concerning the user’s needs or the contextual relevance of the explanation (Warnecke et al. [2020]). It considers the critical aspect of delivering the explanation when it is most needed or can have the most significant impact, ensuring that it remains relevant and beneficial for the user’s immediate context or decision-making process.

Measuring user satisfaction with explanations involves a variety of methods and approaches, as discussed below:

Surveys and Questionnaires – Gathering feedback through structured surveys or questionnaires to evaluate user satisfaction, usefulness, clarity, and other relevant factors.

User Interviews – Conducting interviews to delve deeper into users’ perspectives, allowing for open-ended discussions on their satisfaction levels, concerns, and suggestions.

Post-Interaction Feedback – Prompt users for immediate feedback after interacting with explanations to capture their immediate thoughts and impressions.

User Experience (UX) Metrics – Employing usability tests and metrics (e.g., the System Usability Scale (SUS), Net Promoter Score (NPS), or User Satisfaction Score (USAT)) to observe how users interact with explanations. Noting their ease of use, comprehension, and satisfaction while navigating the system.
Recommendation Systems Metrics – Adapting metrics used in recommendation systems, such as accuracy, coverage, or relevance, to evaluate how well explanations match user preferences or anticipated needs.

Contextual Relevance Assessment – Examining if the explanations are provided at a moment that aligns with the user’s decision-making process or immediate need for information.

3.3 Assessing Trust and Performance

Measures designed to assess human trust in automation focus on two main questions: “Do you trust the AI system’s outputs?” and “Would you follow the AI system’s advice?”. Trust assessment in the XAI context must consider the negative trusting states and whether the user’s trust and reliance on the AI are appropriate. Explanations should let users know whether, when, and why to trust, distrust, or rely on the system. The initially sceptical user may benefit from a good explanation and move into a place of justified trust. However, the subsequent use of the AI system may result in an unexpected outcome that humans would never draw. This surprising event might move the user into an unjustified mistrust, in which the user is sceptical of any outcome the model gives. However, the AI system may provide further explanations, allowing the user to explore the system and converge in a state of appropriate trust and reliance (Hoffman et al. [2018]; Anjomshoae et al. [2021]).

Trusting in XAI systems will always be experimental, so trust and reliance relationships should maintain an appropriate, context-dependent state rather than aiming to achieve and maintain a single stable condition (Omeiza et al. [2021]). Explanations should enable users to mitigate unjustified trusting and mistrusting situations and verify the reasons to take the model’s outcome as accurate. Moreover, explanations should allow users to understand circumstances in which the model’s outcome may not be trustworthy and should not be followed even though the result seems trustworthy (e.g., high confidence score). Assessing if explanations support trust calibration involves evaluating whether the explanations effectively guide users in forming appropriate trust, distrust, or reliance on the AI system. Here are steps to assess trust calibration:

Scenario-Based Assessments – Presenting users with hypothetical scenarios where they must make decisions based on AI-generated information and assessing whether the explanations influenced their decision-making and trust in the AI.

Post-Interaction Interviews – This includes specific questions in surveys or interviews that ask users about their perceived trust levels before and after receiving explanations. For example:

- “Has the explanation influenced your trust in the AI system?”
- “Did the explanation help you understand when to trust or distrust the AI system?”
- “Did the explanation provide enough information for you to decide whether to trust the AI?”
- “Did the explanation cover the factors that affect trust in the AI system?”

Comparison Studies – Comparing users’ perceived trust and decision-making in scenarios where explanations were provided versus scenarios without explanations. Comparative analysis can highlight the impact of explanations on trust calibration.

Longitudinal Studies – Conducting studies over an extended period to assess how users’ trust evolves with consistent exposure to explanations and evaluate changes in trust patterns and decision-making based on the explanations received.

Performance assessment aims to determine the degree of success achieved at conducting a task through human-XAI interaction. The human user’s performance depends on the qualities of their mental model (e.g., correctness, completeness) and the trust placed in the AI system. The performance will improve as a result of being given satisfying explanations that create an accurate mental model and foster trust. The evaluation of the performance is two-fold: the user’s performance and the user-XAI system’s performance (Hoffman et al. [2018]).

3.3.1 User Performance Evaluation

User performance evaluation focuses on assessing the user’s efficacy, understanding, and execution of tasks using XAI systems. User performance is measured by task completion, decision-making accuracy, and adoption rate of AI-generated explanations.
Task Completion – Measuring how successfully users perform tasks with the assistance of XAI explanations. This could involve completion rates or task accuracy before and after receiving explanations.

User Decision-Making – Evaluating the quality of decisions made by users with and without access to XAI explanations and considering metrics like decision accuracy, confidence levels, or time taken to make decisions.

Explanation Adoption Rate – Measuring the rate at which users accept and use the explanations provided by the XAI system and analyzing if users incorporate these explanations into their workflows or decision processes.

3.3.2 User-XAI Interaction Evaluation

The user-XAI system’s performance is measured based on their collaborative ability to achieve shared objectives or tasks. This encompasses a broader evaluation of the dynamic between the user and the XAI system, which evaluates the collaborative performance, communication and error analysis, considering the effectiveness of interactions with the AI system.

Collaborative Performance Metrics – Measuring the efficiency, accuracy, or effectiveness of the collaborative efforts achieved through the interaction, such as collective decision-making accuracy or overall task success.

Communication Analysis – Evaluating the quality and clarity of communication between users and the XAI system during task execution. This could involve analyzing dialogue logs or communication transcripts.

Error Analysis – Evaluating how errors or misunderstandings during the collaborative task are handled by the XAI system and users, examining the system’s response and user adaptation.

4 Conclusion

In conclusion, the article navigated through various evaluation paradigms and offered a range of evaluation metrics, providing insights into the comprehensive assessment of explainable AI methods. This classification and the outlined properties serve as a roadmap that can inspire and guide researchers and practitioners seeking appropriate evaluation methods for new and existing XAI systems. These properties are well-defined and adaptable, making them suitable for accommodating various explainability algorithms. By leveraging both objective and subjective evaluation metrics, practitioners and researchers can establish a holistic framework to assess the effectiveness, user acceptance, and robustness of XAI systems. This integrated approach is anticipated to facilitate a deeper understanding of XAI systems’ capabilities, fostering trust, reliability, and adaptability in the constantly evolving landscape of AI-driven technologies.

Acknowledgments

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation.

Supplementary Material

A user satisfaction questionnaire and open-ended questions can be found at the end of the paper.

References


Winston Maxwell and Bruno Dumas. Meaningful xai based on user-centric design methodology: Combining legal and human-computer interaction (hci) approaches to achieve meaningful algorithmic explainability. *Available at SSRN 4520754*, 2023.


Appendix A

User Satisfaction Questionnaire

When designing a questionnaire to assess user satisfaction with explanations, it is recommended to employ a standardized rating system such as the Likert scale, which allows respondents to rate their agreement or satisfaction with different aspects on a scale (e.g., 1 to 5). The questionnaire should be customized based on the context of the explanations and the user group being surveyed.

**Usefulness**
How beneficial was the explanation in aiding your understanding or decision-making?

<table>
<thead>
<tr>
<th>Not beneficial at all</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Extremely beneficial</th>
</tr>
</thead>
</table>

**Completeness**
To what extent did the explanation cover all the necessary information?

<table>
<thead>
<tr>
<th>Incomplete</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Completely thorough</th>
</tr>
</thead>
</table>

**Actionability**
Did the explanation provide actionable insights or guidance?

<table>
<thead>
<tr>
<th>Not actionable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Highly actionable</th>
</tr>
</thead>
</table>

**Relevance**
How relevant was the explanation to your specific query or context?

<table>
<thead>
<tr>
<th>Not relevant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Highly relevant</th>
</tr>
</thead>
</table>
Exploring Evaluation Methodologies for Explainable AI

**Personalization**
Did the explanation feel tailored to your needs or preferences?

Not personalized 1 2 3 4 5 Highly personalized

**Timeliness**
How timely was the delivery of the explanation?

Not timely 1 2 3 4 5 Timely

**Clarity and Understandability**
How clear and understandable was the explanation?

Unclear 1 2 3 4 5 Very clear

**Trustworthiness**
To what extent did you trust the information provided in the explanation?

Not trustworthy 1 2 3 4 5 Very trustworthy

**Ease of Use**
How easy was it to access and interact with the explanation?

Difficult 1 2 3 4 5 Very easy

**Overall Satisfaction**
Overall, how satisfied are you with the explanations provided?

Very dissatisfied 1 2 3 4 5 Very satisfied
Appendix B

Open-ended Questions

Open-ended questions aim to gather detailed feedback and insights from users, providing a deeper understanding of their experiences and preferences regarding the explanations they receive. Open-ended questions are often used along with questionnaires to provide a more comprehensive view of user opinions. Here are some example questions:

1. What aspects of the explanation did you find most useful, and why?
   ...................................................................................................
   ...................................................................................................

2. Can you describe an instance where you felt the explanation lacked important information or context?
   ...................................................................................................
   ...................................................................................................

3. How did the explanation influence your decision-making process or understanding of the situation?
   ...................................................................................................
   ...................................................................................................

4. Were there any specific parts of the explanation that were particularly unclear or difficult to comprehend?
   ...................................................................................................
   ...................................................................................................

5. In what ways do you feel the explanation could have been more tailored to your specific needs or preferences?
   ...................................................................................................
   ...................................................................................................

6. Did the timing of the explanation align well with your expectations or requirements? Please elaborate.
   ...................................................................................................
   ...................................................................................................
Exploring Evaluation Methodologies for Explainable AI

7. What improvements or additional features would you suggest to enhance the usefulness of the explanations?

................................................................................................................................................
................................................................................................................................................

8. Describe an instance where you felt the explanation provided significant value or positively impacted your interaction with the system.

................................................................................................................................................
................................................................................................................................................

9. How did the explanation contribute to building trust or confidence in the information provided?

................................................................................................................................................
................................................................................................................................................

10. Can you share any suggestions or ideas for making the explanations more user-friendly and accessible?

................................................................................................................................................
................................................................................................................................................