A Survey of Evaluation Methods and Metrics for Explanations in Human–Robot Interaction (HRI)

Lennart Wachowiak[1], Oya Celiktutan[1], Andrew Coles[1], Gerard Canal[1]

Abstract— **The crucial role of explanations in making AI safe and trustworthy was not only recognized by the machine learning community but also by roboticists and human–robot interaction researchers. A robot that can explain its actions is supposed to be better perceived by the user, be more reliable, and seem more trustworthy. In collaborative scenarios, explanations are often expected to even improve the team's performance. To test whether a developed explanation-related ability meets these promises, it is essential to rigorously evaluate them. Due to the many aspects of explanations that can be evaluated, and their varying importance in different circumstances, a plethora of evaluation methods are available. In this survey, we provide a comprehensive overview of such methods while discussing features and considerations unique to explanations given during human–robot interactions.**

I. Introduction

As the deployment of AI technology increases, the need for AI to be reliable, transparent, and trustworthy becomes a core concern for research, industry, and society at large. Explanations are one of the mechanisms to ensure this need is met. This also holds true for robotic AI, where explanations will allow humans to trust their robot collaborators and to safely rely on them. To ensure that explanations meet their proposed goals, we need to be able to evaluate developed explainability mechanisms.

In this paper, we look at various evaluation methods and metrics used in the field of explainable AI (XAI), summarize findings, and evaluate their relevance in the context of human–robot interaction (HRI). We distinguish between methods that evaluate:

1) the content quality of an explanation (Sec. II),
2) the effect of an explanation (Sec. III),
3) the faithfulness of an explanation (Sec. IV),
4) and the timing and need for an explanation (Sec. V).

We compile an extensive list of available evaluation metrics and set our paper apart from existing surveys on explanation evaluation (such as [1], [2], [3], [4], [5], [6]) by providing an HRI perspective. We, moreover, aim to showcase a diverse set of evaluation metrics, giving space to possibilities such as participatory design and think-aloud protocols that found little attention elsewhere.

II. Evaluating the Content Quality of an Explanation

A core aspect of XAI evaluation is evaluating the content of an explanation directly by checking it against criteria that test for some desired qualities. Such qualities are, for instance, the redundancy of information in an explanation or the explanation's intelligibility. Whether an explanation adheres to these criteria can be either assessed by the researchers themselves using checklists (Sec. II-A), a quantifying metric (Sec. II-B and II-D), or by potential users of the AI system via interviews and questionnaires (Sec. II-C) or participatory design (Sec. II-E).

A. Checklists for Researchers

Checklists can be used to a priori evaluate the quality of explanations without the need for user involvement. The researchers themselves take the explanations generated by their agent and compare them against the checklist. For instance, Hoffman et al. [1] propose a checklist that asks the researcher whether an explanation: improves the user's understanding, is satisfying, sufficiently detailed, sufficiently complete, helps with using the software, and shows whether the software is accurate, reliable, and trustworthy. However, given that the researcher makes these judgments themselves, the answers might only be crude approximations of the explanation's real quality.

B. Metric Scores

As an alternative to explanation quality criteria being subjectively evaluated by the researcher, some XAI researchers evaluate their explanations with metrics that can be mostly automatically computed based on the generated explanations or model implementation. While these metrics are often said to be objective evaluations, the selection of which criterion to use, how to parameterize it, and how to weigh it in the explanation's overall assessment is still very subjective.

Examples of such metrics are given by Rosenfeld [7], who suggests metrics that make it possible to quantitatively evaluate an explanation and compare it to others. One of the suggested metrics adds a penalty to a system's performance based on how many rules are used in an explanation. This metric is, for example, applicable to decision trees where the number of rules is easily countable. Similarly to this metric, the second one penalizes the number of input features processed by a model, based on the assumption that a model and an explanation using fewer input features is easier to understand. Another metric measures the explanation's stability given slightly perturbed inputs.

*This work was supported by UKRI (EP/S023356/1), in the CDT in Safe and Trusted AI, the CHIST-ERA project COHERENT (EP/V062506/1) and the EPSRC project LISI (EP/V010875/1). We also thank Toyota Motor Europe and Toyota Motor Corporation. G. Canal was supported by the RAEng and the Office of the Chief Science Adviser for National Security under the UK IC Postdoctoral Research Fellowship programme.

[1] All authors are with King's College London, London, UK, WC2R 2LS. Contact: lennart.wachowiak@kcl.ac.uk, https://lwachowiak.github.io/

Furthermore, Vilone and Longo [2] provide an extensive list of individual studies making use of different metrics scores that can be evaluated without relying on the users' or researchers' opinions. Among others, they find examples of metrics scoring the XAI method's sensitivity to changes in input or the parameters of the explained model, the explanation's completeness, and text quality.

C. *User Perspectives*

Instead of leaving the evaluation to the researchers themselves, it is common to let users judge an explanation's content. This can be done via interviews or questionnaires, which often let the users judge different criteria via scales. Participants for such user studies can be sampled randomly or from the pool of people representing the future end users of the XAI agent.

An example of a scale for measuring an explanation's quality is Holzinger et al.'s System Causability Scale [8]. It consists of 10 items the user has to rate on a Likert scale [9]. Among others, the questionnaire asks if the explanation was understandable, if its level of detail could be adapted, and if it was given in a timely manner. Furthermore, Hoffman et al. [1] provide a version of their checklist for researchers, mentioned in the previous section, that is adapted for users. Silva et al. [10] provide another questionnaire that they tested in a large human–agent interaction study with 286 participants, also showing its correlation with trust and performance measures. The questionnaire consists of 30 questions that target the three axes of simulatability (e.g., "I would not understand how to apply the explanations to new questions."), transparency (e.g., "I understand why the agent used specific information in its explanation."), and usability (e.g., "The explanations were useful."). Lastly, Zemla et al. [11] investigate what criteria of an explanation correlate with people's judgment of an explanation's goodness. They find, for instance, that people prefer complex explanations that refer to multiple causes, and they identify criteria such as being well articulated and internally coherent as predictors of explanation quality. The finding on complex explanations is opposed to the idea of preferring simpler explanations, which is, for instance, supported by Rosenfeld's metrics [7]. Findings like this can inform the design of future explanation quality checklists and questionnaires. Regarding the issue of explanation complexity, we can see that it seems to be domain- and context-dependent, and we should not make a universal judgment as to whether simpler or more complex explanations are preferable.

Instead of using a scale or questionnaire, it is possible to put a stronger focus on qualitative data. For example, Spinner et al. [12] opt for free-form answers by conducting semi-structured interviews with different potential user groups of an XAI tool they developed. Additionally, they also collect user impressions from think-aloud sessions in which the users communicate their thoughts while they try out the tool. The potential use of think-aloud sessions [13], [14] is further discussed in Section V-B.

While questionnaires and interviews are the most common ways to consider users' perspectives, there exists research that uses indirect measures derived from social or physiological cues. For instance, Guerdan et al. relate participants' facial affect during an interaction to their use and impression of an explanation [15]. Similar analyses could be envisioned for social and physiological cues of other modalities. Such analyses have the potential to be helpful in long-term studies of explanations in real-world applications where you do not want to ask the user repeatedly for feedback. However, such analyses also require a complicated data monitoring and collection setup, an analysis pipeline, and data handling considerations. Moreover, it is much harder to reliably understand how the collected cues relate to a person's opinion on an explanation compared to simply asking them via questionnaires or interviews.

D. *Matching Human Explanations*

Another form of evaluating the quality of explanations is comparing them to a human-generated gold standard. Instead of asking people for their opinion, one can directly compare the agent's explanation with a human explanation collected beforehand. This type of evaluation does not take faithfulness, discussed in Section IV, into account but only checks if an explanation is similar to that of a human. In other words, only because an agent's explanation sounds like that of a human, it is not guaranteed that the explanation takes into account the model's true inner decision-making process. The advantage of this type of evaluation is that evaluation is fast, and different explanation generation methods can be compared against the same collected gold standard of human explanations. Usually, the average similarity can be expressed as a numeric score, thus, making it usable as a metric score. However, a human-generated dataset must be available, which is rarely the case and is expensive to collect.

Such evaluation techniques have been employed in computer vision [16], where human annotators created explanations by annotating areas in an image that were relevant to its classification, and in natural language processing [17], [16], where human annotators can provide their rationale for a text classification by saying which words of the input are important. Ehsan et al. [18] go beyond simply using human explanations for comparison and directly train an agent to mimic human rationales given state–action pairs of the game Frogger and human-provided explanations. In the future, we assume that models such as GPT-4 [19] that are good at imitating styles and can process multi-modal input data will lead to further advances in generating such human-like explanations.

E. *Participatory Design*

Instead of only asking for the user's perception of an explainable agent once it has been developed, participatory design can be used to co-create an explanation that respects the end user's needs. Through participatory design methods, users are involved from the beginning and provide feedback before the concrete XAI method is fully developed. This way, a designed XAI mechanism is evaluated not only in the end but throughout the whole development process. An example of participatory design is a workshop with domain experts that indicate their potential explanation needs, describe the real world use-cases, and are asked to actively provide design suggestions. The use of participatory design for XAI has been proposed several times [20], [21], [22]. The most concrete suggestion, proposed by Eiband et al. [21], was a participatory design process consisting of five stages. The first three stages aim to identify the optimal content for the to-be-developed explanations for illuminating an opaque system. Therefore, the researchers need to identify (1) what actually happens inside the system's architecture and what can be explained, (2) what the user thinks happens, and (3) which aspects need explanation according to the users. The last two stages are concerned with the presentation format and how the necessary changes in the users' mental models can be achieved through explanations.

F. Role in HRI

Evaluating an explanation using checklists for researchers and questionnaires targeted at the end user is highly relevant in the HRI context. However, many of the existing questionnaires are developed with non-embodied machine learning models in mind that, for example, classify medical images or make loan decisions. Thus, there is a need for questionnaires more attuned to the HRI context, tackling the specific issues that arise when a robot provides an explanation to a human. For instance, a robot needs to integrate the process of giving an explanation naturally into the interaction. Content delivery should usually be polite and non-interruptive but still get the user's attention. Moreover, a robot is embodied and, therefore, can use its body to communicate an explanation. For instance, a humanoid robot can reinforce its point with gestures or point at something in the environment. A non-humanoid robot, on the other hand, might use flashing lights as an additional mode of communication. As the whole interaction is embedded into an environment, it is also possible to envision using augmented reality systems that highlight parts of the environment to the user that are important to the explanation.

III. Evaluating the Effects of an Explanation

Besides evaluating the explanation content, it is usually insightful to evaluate the downstream effects of giving an explanation. These effects can be categorized into three groups: changes in the performance of the human–agent team, changes in how the user perceives the agent, and changes in the user's mental model of the agent and task.

A. Performance

Explanations can strongly impact the performance of a human–robot team. In many scenarios, one expects explanations to cause a performance increase, given that the user better understands the robot and, thus, can better anticipate its actions and adapt to it. On the other hand, performance might decrease due to added cognitive load for the user or the use of less powerful but more explainable AI models. To understand the performance impact of an explanation, one can employ task-specific performance measures. For instance, one can measure the time needed to complete a task, the number of correct decisions made by the human–agent team, the number of errors avoided, or the number of fulfilled sub-tasks per session. Different performance measures can provide opposing insights. For example, an explanation might be given as additional safety rail to ensure a system makes the right decision, and thus additional time for approving the agent's decisions is expected. While the performance measured in time might worsen in this case, the performance measured in the number of prevented errors should improve. Moreover, in some cases, performance can also be separately analyzed for the user, the agent, and the human–agent team.

Furthermore, when evaluating explainable black-box models, Rosenfeld [7] suggested a metric measuring the difference in the accuracy between a proposed black-box model to an alternative transparent model. The idea behind this metric is that black-box models should only be used if they perform significantly better than a transparent alternative. In the case of robots, this metric could be applied, among others, to vision modules, where the use of difficult-to-understand neural models is commonplace.

B. Perception of Robot

A common assumption is that a robot that can explain itself will be better perceived by users than one that can not. A core construct measured in this context is the user's trust in the agent, with much of XAI research citing increasing the user's trust as motivation. This is the case since explanations can show the user how an agent works, that it makes decisions based on the right reasons, and that its future behavior is predictable. At the same time, explanations can also decrease trust. For example, if the agent makes a correct decision but the explanation shows that it was made for the wrong reasons, the user should lower their trust level. Thus, explanations allow the user to calibrate their trust appropriately. Trust can be measured via the user's behavior with regard to the agent. For example, one can test whether users are more likely to follow the suggestions of an agent if it can give explanations. Similarly, it can be informative to measure how much the user interacts with the agent and if explanations increase the number of interactions [3].

More commonly, trust is measured via questionnaires, some of which we present in the following. Hoffman et al. [1] provide a questionnaire with eight Likert scale questions the user has to answer. Items include statements about the user's confidence and wariness as well as the model's predictability, efficiency, and reliability. Hoffman et al.'s scale is developed for explainable AI models and needs rephrasing to apply to a robotics scenario. Overall, there exist many trust scales targeting different contexts, for instance, trust in industrial human–robot collaboration [23] or human–robot interaction [24], [25] — however, these scales might not always target aspects of the robot that can be changed through explanations. It can be hard to choose the right scale for a given context, and existing scales often do not fit a researcher's experiment perfectly, with some items not applying to the new context [26]. The available variety of scales makes it also more difficult to make comparisons across studies. Lastly, literature distinguishes between emotional and cognitive trust [27], [28] with emotional trust being based on affect rather than rational thinking.

Beyond trust, there exists a plethora of dimensions across which the user's perception of the robot can be measured. Possibilities include persuasiveness [29], likeability, anthropomorphism, or perceived intelligence [30]. Which of these a researcher chooses to evaluate depends on the context in which the explanation is given and on the explanation's goals.

C. Investigating the User's Mental Model

Researchers can check users' understanding of an agent and the interaction by probing them with specific questions. Before investigating the user's understanding of the agent, the user needs to observe the AI for a while or interact with it. Afterwards, the researcher shows the user a new scenario in which the agent is required to make a decision. Then, the researchers ask the user to predict how the agent would act in such a scenario, thereby, testing whether the user gathered an appropriate understanding of the agent's internal mechanisms. The researcher would then compare if agents that explained their decisions in the first phase led to more correct and complete mental models of the users. Doshi-Velez and Kim [4] distinguish between forward simulation, where the user has to predict the model's output given an input, and counterfactual simulation, where the user is given a model's input, its output, and an explanation for its decision and then has to tell what changes are needed for the model output to change to the correct output.

Another method for probing a user's mental model comes from situation awareness research and human factors studies. Sanneman and Shah [31] propose using the situation awareness-based global assessment technique [32] that tests the user's informational needs at various points throughout the interactions via a set of questions. With appropriate XAI techniques, these needs should be met at each point in time.

D. Role in HRI

Studying the effects an explanation has on the interaction and the user is essential in HRI. As in general XAI research, increasing trust is an often cited motivation [33]. Building trust might be especially relevant when the user encounters a robot compared to a virtual AI as a literature review comparing the two cases found that users initially start out with low trust in a robot and increase their trust after gathering experience with it [27]. The typical trust trajectory with virtual AI often develops in the opposite way, starting out high and lowering with time. Thus, explanations can be a way of letting the user overcome their initial distrust or even fear of the robot that otherwise might prevent them from interacting with the robot altogether. Secondly, task performance can be a core goal in human–agent collaborations, for instance, when robots and humans assemble a product together in manufacturing. Lastly, investigating the user's mental model can be applied as well. It is important that the user understands what the robot will do, but also how the environment will develop, and what the best team-level decisions are given the task. These different themes of understanding can all be supported by explanations.

IV. Evaluating the Faithfulness of an Explanation

Even if an explanation is perfectly formulated and convinces the user, it does not necessarily follow that this explanation captures the actual reasons the model based its decision on. To generate appropriate trust in an agent instead of nurturing false expectations and unsafe reliance on AI, it is important to provide explanations that reflect the AI's true decision-making process. However, explanation generation methods vary in their degree of faithfulness, with model-agnostic post-hoc explanation methods often just estimating the actual inner workings of a model. Another reason for giving less faithful explanations can be to reduce the explanation's length. For instance, in explainable planning, one might want to only refer to a limited number of preconditions, actions, and goals to not overload the user with information. Zhou et al. [6] discuss faithfulness using the term fidelity, focusing not only on whether an explanation is true to the underlying decision-making mechanics but also on whether it is complete.

Faithfulness can not be evaluated based on user perception or performance metrics [34]. Therefore, researchers have proposed different ways to estimate a model's faithfulness. For example, Dasgupta et al. [35] measure consistency and sufficiency when evaluating the faithfulness of black box prediction models. For consistency, they quantify how often two inputs with the same explanation receive the same output label from the model. For sufficiency, they measure that if a system explains a prediction it made, other inputs get the same prediction if one can apply the same explanation.

A. Role in HRI

The importance of faithfulness depends on what robotic ability is explained and how that ability is implemented. For instance, in explainable planning [36] an explanation can use the same symbolic representations that the planner itself uses and straightforwardly compare costs between different potential plans. Instead of dealing with issues of faithfulness, explainable planning deals with issues of making explanations easily understandable by humans [37] and finding the right level of abstraction. If the planner provides non-optimal plans explaining faithfully potentially becomes more difficult as the explanation can not argue for an optimal solution. Instead, the explanation needs to say why the planner did not arrive at the optimal solution or why the provided plan is still sufficient. However, a robot might also want to explain why it made a certain object classification or why it understood a language utterance the way it did. For neural models used in such tasks, the faithfulness of an explanation is an important point of scientific inquiry [34].

V. Evaluating Timing and Need for an Explanation

While evaluating an explanation's quality and effect is important, in HRI it is also essential to consider when explanations are given. A given explanation can be understandable, faithful, and well formulated, but be given at a time when the user has no need for it — thus, being distracting instead of helpful. To some extent, evaluating the effects of an explanation already targets the issue of timing and need. For example, an agent that always gives explanations although they are not needed might annoy the user and cause additional cognitive load, thereby causing a worse performance. However, there are additional evaluation methods available that more precisely target the issue of timing and relevance. Firstly, these methods might target interactions in which explanations are already given to check whether those explanations are being provided at the right time. Secondly, they might be employed before explanations are developed to identify what needs explaining and at what time.

A. Questionnaires

Checklists and questionnaires can help determine if someone needs an explanation at a given moment. For example, one can employ methods from situation awareness research [31] that try to identify if a person's information need is met at a given point in time. At times when the user is not sufficiently aware of the agent's functionality or the overall task, an explanation might be helpful. On the other hand, if one evaluates a scenario where the agent already provides explanations, one can question the user directly after an explanation. These questions can then help us understand to what extent the explanation was actually needed, why the user requested it, or if it was helpful even without having been actively requested by the user. In the case of user-requested explanations, questionnaires such as the curiosity checklist by Hoffman et al. [1] can be used. The curiosity checklist asks about what the user wants to know from the agent and why they have asked for an explanation.

Lastly, we want to mention the findings from Liquin and Lombrozo [38] as they identify indicators of the need for an explanation that could be used to better evaluate the relevance of a given explanation in future assessment methods. They empirically test a set of 13 candidate indicators of the need for an explanation, for instance, the expected future utility of receiving the explanation. To do so, participants were given why-questions from Reddit and were asked how strongly these demand explanations. Additionally, participants had to rate the 13 potential need indicators. Besides expected utility, requiring expertise to answer the question was a strong predictor of the need for an explanation. Other valuable predictors included the user's prior knowledge of the topic and the explanation's expected information content.

B. Intropsective User Reports

To understand the user's internal mental state and identify moments of curiosity, confusion, and those in which the need for an explanation arises, one can use think-aloud methods [14] or interviews. The think-aloud method makes the user narrate their thought process throughout an interaction (concurrent think-aloud) or while re-watching the interaction (retrospective think-aloud). Afterwards, the researcher can analyze the collected verbal reports. There are various aspects to consider, for example, a concurrent think-aloud session might change how the user interacts with the system, while a retrospective think-aloud session might be less true to the original thought process of the user. It is, moreover, important to allow the user to get used to the method via some initial exercise, not to let the session run too long, and to only let participants narrate their thoughts instead of meta-explanations about their thought [13]. Secondly, there exist specific interview methods that allow for a deep dive into a person's phenomenological experience and provide more qualitative descriptions, such as the micro-phenomenological interview, in which the interviewee re-immerses themselves in their past experience, trying to relive it [39]. Another example is the descriptive experience sampling method [40], where people are prompted via a signal from a beeper going off in their natural, non-lab environment and then are asked to write down notes about their inner experience at the moment before the beep. However, these phenomenological methods have not yet been tested in the context of exploring someone's explanation needs.

Some research uses these methods to identify moments in which the need for an explanation arises in scenarios where explanations were not yet provided to the user. For example, Wachowiak et al. [41] use retrospective think-aloud interviews to get insights into the thoughts and mental states of users interacting with an agent in a collaborative cooking game. Based on the verbal reports, two annotators annotated points in time during which the user was confused or the agent made errors, that is, scenarios where an explanation would have been helpful. The retrospective think-aloud protocol can also be used in scenarios where explanations were already given. It can help to understand the user's explanation need while not influencing the interaction as it happens.

C. Role in HRI

The timing of explanations and identifying what caused the need for an explanation seems more critical in HRI than in other fields. In the simple case of an image classification model whose output is checked by a user, it might suffice to have a button with which the user can request an explanation. In an interactive human–robot collaboration, however, this will not suffice. For example, the user might be focused on their own sub-tasks for which they might not ask the robot for an explanation, although it might still be helpful for improved team coordination or preventing wrong decisions by the robot. In another scenario, the robot might not be equipped with natural language understanding capabilities and unable to answer explicit explanation requests, thus, needing to identify by itself when an explanation is needed. Similarly, an explicit request might be impossible when dealing with a person with a disability.

Furthermore, an HRI has many facets — the task itself, the environment, and the robot — each being potentially the source of a user's explanation need. In addition, many robots have a diverse set of capabilities, including natural language understanding, perception, object manipulation, and high-level action planning. All these facets are part of HRI and might demand explanations — thus, it is crucial to understand what explanations are needed and when they are needed. Methods presented in this section can help with that, as well as user-oriented research methods such as those from participatory design presented in Section II-E.

VI. Discussion and Conclusion

In this paper, we presented a variety of evaluation methods that can be used to assess explainable agents and robots.

Based on our survey and knowledge of the field we want to end the paper with the following recommendations for evaluating explanations in HRI:

1) **Identify what aspect of the explanation you want to evaluate. Then, state clearly what you evaluate.** Instead of saying that you evaluate your explanation, say that you evaluate your explanation's effect on performance or your explanation's faithfulness.
2) **Employ multiple methods.** There is no single standardized XAI evaluation method that can always be used. Instead, the presented evaluation methods have different strengths and weaknesses. What is most informative depends on the application, users, available resources, and the goal of the explanation. Different evaluation methods can tell different stories [42], [43]. Thus, it is advisable to employ multiple methods and compare outcomes.
3) **Consider the context and goal of an explanation.** Not all criteria for explanation quality are important in every situation. For instance, research has shown that in some contexts, more complex explanations are preferred over simple explanations [11], which, however, would be penalized by some metrics, e.g., [7].
4) **Do not rely solely on pre-made checklists.** This rule directly follows from the point made before. Individual checklist criteria are often situation dependent. To know which criteria are important, one needs to acquire an in-depth understanding of how explanations are used in practice. This usually means involving users.
5) **Consider the types of explanation recipient.** When considering the context in which an explanation is given, it is important to look not only at the task and what the explanation is used for but also at the different possible explanation recipients. Different user groups, such as end users, developers, or external entities, might require different explanations [44]. Similarly, users with different socio-cultural backgrounds might have different explanation preferences [45]. How to adapt to the explanation process appropriately is still an open problem. To fully understand how your explanations are perceived, it is necessary to carefully consider with whom to evaluate and make it explicit in the reporting.
6) **Benchmark your explanation against other explanations.** Consider already existing methods of providing explanations when evaluating your newly developed XAI methods and benchmark against them. Comparable approaches might not always be available as explanations are often developed for very context-specific cases.
7) **Be scientifically rigorous.** The field of HRI is likely not spared from reproducibility and reliability problems that hit other fields of behavioral study [46]. Thus, it is of vital importance to show statistical and methodological rigor in evaluating explanations. That means, for example, pre-registering your hypotheses and evaluation protocol, justifying sample sizes via a power analysis when evaluating with users, and controlling for multiple comparisons.

References

[1] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance," *Frontiers in Computer Science*, 2023.

[2] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, 2021.

[3] J. H.-w. Hsiao, H. H. T. Ngai, L. Qiu, Y. Yang, and C. C. Cao, "Roadmap of designing cognitive metrics for explainable artificial intelligence (xai)," *arXiv preprint arXiv:2108.01737*, 2021.

[4] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[5] L. Coroama and A. Groza, "Evaluation Metrics in Explainable Artificial Intelligence (XAI)," in *ARTIIS*, Springer, 2022.

[6] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, 2021.

[7] A. Rosenfeld, "Better metrics for evaluating explainable artificial intelligence," in *AAMAS*, 2021.

[8] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations," *KI-Künstliche Intelligenz*, 2020.

[9] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, 1932.

[10] A. Silva, M. Schrum, E. Hedlund-Botti, N. Gopalan, and M. Gombolay, "Explainable artificial intelligence: Evaluating the objective and subjective impacts of XAI on human-agent interaction," *International Journal of HCI*, 2022.

[11] J. C. Zemla, S. Sloman, C. Bechlivanidis, and D. A. Lagnado, "Evaluating everyday explanations," *Psychonomic Bulletin & Review*, 2017.

[12] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady, "explainer: A visual analytics framework for interactive and explainable machine learning," *TVCG*, 2019.

[13] D. W. Eccles and G. Arsal, "The think aloud method: what is it and how do I use it?," *Qualitative Research in Sport, Exercise and Health*, 2017.

[14] M. Van Someren, Y. F. Barnard, and J. Sandberg, "The think aloud method: a practical approach to modelling cognitive," *AcademicPress*, 1994.

[15] L. Guerdan, A. Raymond, and H. Gunes, "Toward affective XAI: facial affect analysis for understanding explainable human-AI interactions," in *ICCV*, 2021.

[16] S. Mohseni, J. E. Block, and E. D. Ragan, "A human-grounded evaluation benchmark for local explanations of machine learning," *arXiv preprint arXiv:1801.05075*, 2018.

[17] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "A Diagnostic Study of Explainability Techniques for Text Classification," in *EMNLP*, 2020.

[18] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: a technique for explainable AI and its effects on human perceptions," in *IUI*, 2019.

[19] OpenAI, "Gpt-4 technical report," 2023.

[20] H. Mucha, S. Robert, R. Breitschwerdt, and M. Fellmann, "Towards participatory design spaces for explainable ai interfaces in expert domains," in *German Conference on AI*, 2020.

[21] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann, "Bringing transparency design into practice," in *IUI*, 2018.

[22] U. Ehsan and M. O. Riedl, "Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach," in *HCI International 2020 - Late Breaking Papers*, Springer-Verlag, 2020.

[23] G. Charalambous, S. Fletcher, and P. Webb, "The development of a scale to evaluate trust in industrial human-robot collaboration," *International Journal of Social Robotics*, 2016.

[24] R. E. Yagoda and D. J. Gillan, "You want me to trust a ROBOT? The development of a human–robot interaction trust scale," *International Journal of Social Robotics*, 2012.

[25] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the "trust perception scale-HRI"," in *Robust Intelligence and Trust in Autonomous Systems*, Springer, 2016.

[26] M. Chita-Tegmark, T. Law, N. Rabb, and M. Scheutz, "Can you trust your trust measure?," in *HRI*, 2021.

[27] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, 2020.

[28] D. J. McAllister, "Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations," *Academy of Management Journal*, 1995.

[29] M. Dragoni, I. Donadello, and C. Eccher, "Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice," *Artificial Intelligence in Medicine*, 2020.

[30] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, 2009.

[31] L. Sanneman and J. A. Shah, "A situation awareness-based framework for design and evaluation of explainable AI," in *EXTRAAMAS Workshop*, Springer, 2020.

[32] M. R. Endsley, "Situation awareness global assessment technique (SAGAT)," in *NAECON*, IEEE, 1988.

[33] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *AAMAS*, International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[34] A. Jacovi and Y. Goldberg, "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?," in *ACL*, 2020.

[35] S. Dasgupta, N. Frost, and M. Moshkovitz, "Framework for evaluating faithfulness of local explanations," in *International Conference on Machine Learning*, pp. 4794–4815, PMLR, 2022.

[36] M. Fox, D. Long, and D. Magazzeni, "Explainable planning," in *IJCAI-17 Workshop on Explainable Planning*, 2017.

[37] G. Canal, S. Krivić, P. Luff, and A. Coles, "PlanVerb: Domain-Independent Verbalization and Summary of Task Plans," in *AAAI*, 2022.

[38] E. Liquin and T. Lombrozo, "Determinants and Consequences of the Need for Explanation," in *CogSci*, 2018.

[39] C. Petitmengin, A. Remillieux, and C. Valenzuela-Moguillansky, "Discovering the structures of lived experience: Towards a micro-phenomenological analysis method," *Phenomenology and the Cognitive Sciences*, 2019.

[40] R. T. Hurlburt and S. A. Akhter, "The descriptive experience sampling method," *Phenomenology and the Cognitive Sciences*, 2006.

[41] L. Wachowiak, P. Tisnikar, G. Canal, A. Coles, M. Leonetti, and O. Celiktutan, "Analysing eye gaze patterns during confusion and errors in human–agent collaborations," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2022.

[42] F. Biessmann and D. Refiano, "Quality metrics for transparent machine learning with and without humans in the loop are not correlated," *arXiv preprint arXiv:2107.02033*, 2021.

[43] P. Hase and M. Bansal, "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?," in *ACL*, 2020.

[44] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *AAMAS*, 2019.

[45] H. Kopecka and J. Such, "Explainable AI for Cultural Minds," in *Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction*, 2020.

[46] B. Leichtmann, V. Nitsch, and M. Mara, "Crisis ahead? why human-robot interaction user studies may have replicability problems and directions for improvement," *Frontiers in Robotics and AI*, 2022.