A Meta Survey of Quality Evaluation Criteria in Explanation Methods

Helena Löfström1,2,3, Karl Hammar2, and Ulf Johansson2

1 Department of Information Technology, University of Borås, Borås, Sweden
2 Department of Computing, Jönköping University, Jönköping, Sweden
3 Jönköping International Business School, Jönköping, Sweden
{helena.lofstrom,karl.hammar,ulf.johansson}@ju.se

Abstract. The evaluation of explanation methods has become a significant issue in explainable artificial intelligence (XAI) due to the recent surge of opaque AI models in decision support systems (DSS). Explanations are essential for bias detection and control of uncertainty since most accurate AI models are opaque with low transparency and comprehensibility. There are numerous criteria to choose from when evaluating explanation method quality. However, since existing criteria focus on evaluating single explanation methods, it is not obvious how to compare the quality of different methods.

In this paper, we have conducted a semi-systematic meta-survey over fifteen literature surveys covering the evaluation of explainability to identify existing criteria usable for comparative evaluations of explanation methods.

The main contribution in the paper is the suggestion to use appropriate trust as a criterion to measure the outcome of the subjective evaluation criteria and consequently make comparative evaluations possible. We also present a model of explanation quality aspects. In the model, criteria with similar definitions are grouped and related to three identified aspects of quality; model, explanation, and user. We also notice four commonly accepted criteria (groups) in the literature, covering all aspects of explanation quality: Performance, appropriate trust, explanation satisfaction, and fidelity. We suggest the model be used as a chart for comparative evaluations to create more generalisable research in explanation quality.

Keywords: Explanation method · Evaluation metric · Explainable artificial intelligence · Evaluation of explainability · Comparative evaluations

1 Introduction

AI model-based Decision support systems (DSS) have become increasingly popular due to their possibility of solving a variety of tasks, such as music recommendations or medical diagnosis. However, the highly accurate AI models lack

This research is partly founded by the Swedish Knowledge Foundation through the Industrial Research School INSIDR.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
J. De Weerdt and A. Polyvyanyy (Eds.): CAiSE Forum 2022, LNBIP 452, pp. 55–63, 2022.
https://doi.org/10.1007/978-3-031-07481-3_7
both transparency and comprehensibility in their predictions, which has caused a new field to emerge; *explainable artificial intelligence* (XAI). In this field, the goal is to explain the opaque AI models with the help of explanation methods. However, the interest in explainability has simultaneously led to confusion on numerous fronts; it is, e.g., not clear how to evaluate and compare the quality of explanation methods [1].

The goal of explanations is to strengthen the user in making high-quality decisions by identifying when to trust and not trust the predictions, i.e., to make the user trust the system appropriately. However, many studies that evaluate explanation methods focus on how users experience explanations rather than the degree to which those explanations guide good decision-making. Furthermore, there are no generally accepted criteria for comparative evaluations. Instead, it is up to the individual researcher to decide which criterion to use. At the same time, there are many criteria to choose among, which makes it challenging to compare results from different evaluations. The main research objective in this study was to investigate if there exist generally accepted evaluation criteria, with a well-established method, that is possible to use for comparative evaluations of explanation methods.

A semi-systematic meta-survey was applied to fulfil the research objective. To ensure the quality of our research and find all related papers, we followed the methodology according to [2] but also took into consideration [3]. The study included fifteen literature surveys, covering a theoretical maximum of more than 2700 research articles. The methodology in the paper is divided into three steps: **Designing** the review (including choice of search terms, databases, inclusion and exclusion criteria, which type of information to extract, and the type of analysis), **conducting the review** with documentation of the process, and **analysing** the results based on the choices made in the design. The literature was collected from the areas of Computer Science, Social Science, Business and Marketing Science as well as Decision Science and limited to evaluation criteria for explanation methods. When the final sample was selected, the surveys were analysed to find evaluation criteria (definitions, usage, and quality threshold value). Each evaluation criterion was initially documented separately and then grouped based on definition and aspect of explanation quality. For a detailed description of the methodology in the study, see [4]. Several criteria were subjective and lacked method, which caused challenges in comparing the definitions.

The contributions of this paper are:

- We suggest measuring the outcome of the subjective evaluation criteria and thus make comparative evaluations possible.
- We also present a model, identifying existing evaluation criteria from four different research areas, the aspect of quality they measure and how they relate to each other.

The remainder of this paper is structured as follows: the next section provides a summary of the concept of post hoc explanation methods, while the results and discussion are presented in Sect. 3. The paper ends with some concluding remarks in Sect. 4.
2 Post Hoc Explanations

Research in explanations can be divided into two main focus areas [5,6]; transparency through interpretable models (interpretability) and post-hoc explanation methods for explaining opaque models (explainability). Post-hoc explanation methods apply to the output of the underlying model and create a simplified and interpretable model based on the relation between feature values and the prediction (see Fig. 1).

![Fig. 1. The extraction of Post-hoc explanations, inspired by [7]](image)

The relationships in the interpretable model are presented as explanations and can be, e.g., pixels in pictures, feature importance charts or words in texts, highlighting which features (pixels or words) that are important for the prediction. Explanations are intended to explain the model’s strengths and weaknesses to the users, creating a possibility to identify erroneous predictions and an understanding of the model’s rationale.

3 Results and Discussion

Evaluations of the use of explanation methods could be divided into three aspects of quality; model, explanation, and user (see Fig. 2) [1,8–11]. In line with earlier research, we found that many evaluation criteria had various names in different studies. Based on the findings in the literature, we grouped the criteria on how they were defined: how the criteria were collected, if the criteria were objective (system output) or subjective (human thoughts/output), and what the criteria were supposed to measure. We used the most common or well-established names in the surveys as the name of the criteria (groups). After the analysis, we identified eleven criteria.

3.1 The User Aspect

The majority of the criteria in the user aspect are subjective and challenging to measure. However, since changes in the subjective criteria cause the user’s mental model to change during the usage of the system, we consider the mental model as a container of these criteria and not a criterion of its own [7,10,12–14]. The resulting four criteria connected to the user aspect were:
Fig. 2. High-level model of the criteria groups in the different aspects of explanation quality.

- **Trust**: a highly used criterion and referred to in [14] as one of the absolute most important criteria for success. However, it is inherently subjective and therefore challenging to measure [8,11,14–16]. The definition of trust has been discussed in multiple research disciplines including psychology [17] and machine learning [11]. In [18] trust is related to both credibility and an attitude that an agent will help achieve an individual’s goals in a situation characterised by uncertainty and vulnerability. Trust has, in that way, two aspects: i) the user’s willingness to act based on the recommendation of the system and ii) their confidence in the correctness of the prediction [19]. In earlier research [17,20] two extreme reactions have been studied when users doubt the correctness of a system’s predictions: they put the system aside, not trusting it (*disuse*), or they set their doubts aside and blindly trust the system (*misuse*). There is, in other words, problematic to merely try to increase user’s trust in the system. There must be some form of alignment between the perceived and actual performance of the system.

- **Appropriate trust**: trust is also possible to define as an attitude formed by information about the system and previous experiences. The attitude creates intentions whether or not to rely on, or trust, the system resulting in *reliance* [14,21]. Under this definition, trust is directly connected to the quality of explanations, in the form of *appropriate trust* or *calibrated trust* [7,8,10,11,13–15,22–27]. A good mental model is a requirement for develop-
A Meta Survey of Quality Evaluation Criteria in Explanation Methods

59

The criterion is one of the outcomes of the user’s mental model (see Fig. 2), consisting of the user aspect (human trust in the system) and the explanation aspect (the output of the system) [8, 22]. When identifying the level of appropriate trust, the users try to identify the correct and erroneous predictions. The ideal result is a 1:1 situation where all correct and erroneous predictions are identified (see figure 3). All other cases lead to either misuse (overuse) or disuse (under-use). Appropriate trust could, in other words, be defined as the user’s accuracy; to what degree the users act according to the data.

- Appropriate trust is also closely connected to bias detection which is one of the desired outcomes of explanations. Bias is possible to detect when the user has high appropriate trust and, in that way, can identify the general patterns of errors in the system [14, 24, 26, 27]. However, [24] defines the term as more of a data set desiderata. The criterion could be seen as the joint outcome of the user’s mental model, together with appropriate trust, creating reliance (see Fig. 2).

- Explanation satisfaction: to what extent an explanation user interface or an explanation is suitable for the intended purpose [28]. This criterion is the outcome of the mental model (see Fig. 2). The evaluation is conducted with Likert scales, with questions similar to, e.g., if the user understands the explanations and finds them relevant.

| Actual value | Prediction |
|--------------|------------|
| T: T         | T: F       |
| F: T         | F: F       |

Fig. 3. Appropriate trust is the user’s ability to detect (trust) the true predictions and detect (distrust) the false predictions.

3.2 The Explanation Aspect

The Explanation aspect of quality is placed between the model aspect and the user aspect, focusing on the evaluations of the explanations. It is sometimes referred to as the user-machine performance, highlighting the relation between the machine and the human. The criteria are objective but can include humans in the evaluation. Five criteria were identified in this aspect:

- Fidelity: reflects how accurately the explanation method mirrors the underlying model [6, 7, 10–12, 24–26]. The fidelity could also be used to measure the
difference between description accuracy given by the system and the description accuracy assessed by the user, including a more subjective perspective [24].

- **Identity, Separability, Novelty, and Representativeness:** in [11] the authors suggest several new metrics for evaluation of explanations. The criteria are related to fidelity and catch the explanations’ correctness, although they can be calculated without the inclusion of humans. The three first criteria compare the explanations between different instances: i) identical instances should have identical explanations, ii) non-identical instances should not have identical explanations, and iii) the instance should not come from a region in instance space far from the training data. Finally, representativeness measures how many instances are covered by the explanation.

### 3.3 The Model Aspect

In the explanation aspect, the output from the user aspect, or mental model, is fed back to the system via the criteria, *appropriate trust* and *reliance* (see Fig. 2). Several evaluation criteria are connected to the underlying model’s aspect of quality:

- **Performance:** a group of criteria that, together with appropriate trust, is most often referred to in the surveys as important to the quality of the explanations. The criteria indicates the level of correctness of the model against the actual target, such as *accuracy* and is measured similar to appropriate trust (see Fig. 3) [1,7,10,11,13,14,16,23–26,29].

- **Fairness:** the opposite to bias, i.e., to what degree does the model have general errors patterns. However, they measure the same aspect of quality. A model without bias is unbiased [8], i.e., have a high level of fairness. The level of bias indicates the level of fairness and also the level of reliability of the model (see Fig. 2). The criteria are considered by some authors more of desiderata and are therefore somewhat lacking in method description [11].

- **Reliability:** close to the criterion Accuracy, since it signals a confidence measure of the model to the user in a specific situation [7,8,10,11,23,24]. Reliability is possible to measure in different ways. In, e.g., [10] the authors use questionnaires with likert scales and questions similar to if the XAI system is reliable.

### 3.4 Discussion

Comparative evaluations of explanation methods are challenging, and a general computational benchmark across all possible explanation methods is seen in [16] as unlikely to be possible due to the subjective characteristics. However, it is generally accepted that the mental model affects the level of appropriate trust and reliance in the system. If considering the mental model as a container of the subjective criteria, the outcome of the mental model would be possible to measure through the criteria appropriate trust (see, e.g., [10]). Although appropriate
trust does not explicitly answer how the user experiences the explanations, it demonstrates if they fulfil one of the most crucial goals of explanation methods; if the user can detect correct and erroneous predictions. By measuring the outcome of the mental model through appropriate trust, we get an objective metric for the quality of the user aspect and create possibilities for comparative evaluations of explanation methods.

It is essential to highlight that we do not consider the subjective criteria unnecessary or unimportant to measure. In contrast, we acknowledge them as crucial for quality. When evaluating a single explanation method, it could be vital to follow subjective criteria changes. However, if the evaluation intends to gain comparable results, we recommend using the appropriate trust criteria.

4 Conclusion

This paper conducted a semi-systematic meta-survey over fifteen surveys to find commonly accepted evaluation criteria, with a well-defined method possible to use in comparative evaluations. The field was found to focus on human-in-the-loop evaluations, creating severe challenges for quality comparisons.

The major contribution of the paper is the suggestion of using the criterion appropriate trust as an outcome metric of the subjective criteria in the mental model, overcoming the problems with comparative evaluations. We also present a high-level model of explanation method quality, identifying three aspects of quality: model, explanation, and user. The quality of an explanation method is suggested to be a composition of all three aspects. We also identified ten evaluation criteria (groups) in different quality aspects and their relation in the meta-survey. Four of the criteria were mentioned in more than half of the surveys as necessary for the quality of an explanation method: performance, appropriate trust, explanation satisfaction, and fidelity. These criteria cover all three aspects of explanation quality, and we suggest that they are used when researchers want to make their evaluations comparable.

The mental model consists of several subjective criteria, all of which can affect the output. We suggest as future work evaluations with human respondents studying how the subjective criteria affect the level of appropriate trust in the system.

References

1. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. 116(44), 22071–22080 (2019)
2. Snyder, H.: Literature review as a research methodology: an overview and guidelines. J. Bus. Res. 104, 333–339 (2019)
3. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: writing a literature review. MIS Q. xiii–xxiii (2002)
4. Löfström, H., Hammar, K., Johansson, U.: A meta survey of quality evaluation criteria in explanation methods (2022)
5. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777 (2017)

6. Moradi, M., Samwald, M.: Post-hoc explanation of black-box classifiers using confident itemsets. Expert Syst. Appl. 165, 113941 (2021)

7. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115 (2020)

8. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

9. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89. IEEE (2018)

10. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)

11. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics 8, 832 (2019)

12. Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A., Klein, G.: Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint arXiv:1902.01876 (2019)

13. Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable AI systems. arXiv, pp. arXiv-1811 (2018)

14. Gunning, D., Aha, D.W.: Darpa’s explainable artificial intelligence program. AI Mag 40(2), 44–58 (2019)

15. Hoff, K.A., Bashir, M.: Trust in automation: integrating empirical evidence on factors that influence trust. Hum. Factors 57(3), 407–434 (2015)

16. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics 10(5), 593 (2021)

17. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. Int. J. Hum Comput Stud. 58(6), 697–718 (2003)

18. Pavlidis, M., Mouratidis, H., Islam, S., Kearney, P.: Dealing with trust and control: a meta-model for trustworthy information systems development. In: 2012 Sixth International Conference on Research Challenges in Information Science (RCIS), pp. 1–9. IEEE (2012)

19. Yang, F., Huang, Z., Scholtz, J., Arendt, D.L.: How do visual explanations foster end users’ appropriate trust in machine learning? In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 189–201 (2020)

20. Marsh, S., Dibben, M.R.: Trust, untrust, distrust and mistrust – an exploration of the dark(er) side. In: Herrmann, P., Issarny, V., Shiu, S. (eds.) iTrust 2005. LNCS, vol. 3477, pp. 17–33. Springer, Heidelberg (2005). https://doi.org/10.1007/11429760_2

21. Ekman, F., Johansson, M., Sochor, J.: Creating appropriate trust in automated systems: a framework for HMI design. IEEE Trans. Hum. Mach. Syst. 48(1), 95–101 (2017)

22. McDermott, P.L., Ten Brink, R.N.: Practical guidance for evaluating calibrated trust. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 63, pp. 362–366. SAGE Publications Sage CA, Los Angeles (2019)
23. Chromik, M., Schuessler, M.: A taxonomy for human subject evaluation of black-box explanations in xai. In ExSS-ATEC@ IUI (2020)
24. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv preprint arXiv:2006.11371 (2020)
25. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160 (2018)
26. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable AI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, pp. 1–15, New York, Association for Computing Machinery (2019)
27. Zhang, Y., Chen, X.: Explainable recommendation: a survey and new perspectives. arXiv preprint arXiv:1804.11192 (2018)
28. Holzinger, A., Carrington, A., Müller, H.: Measuring the quality of explanations: the system causability scale (SCS). KI-Künstliche Intelligenz, pp. 1–6 (2020)
29. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy 23(1), 18 (2021)