A Quantitative Human-Grounded Evaluation Process for Explainable Machine Learning

Katharina Beckh*[,1,3], Sebastian Müller*[,2,3] and Stefan Rüping[1]

[1]Fraunhofer IAIS
[2]University of Bonn
[3]ML2R - Competence Center Machine Learning Rhein-Ruhr

Abstract

Methods from explainable machine learning are increasingly applied. However, evaluation of these methods is often anecdotal and not systematic. Prior work has identified properties of explanation quality and we argue that evaluation should be based on them. In this position paper, we provide an evaluation process that follows the idea of property testing. The process acknowledges the central role of the human, yet argues for a quantitative approach for the evaluation. We find that properties can be divided into two groups, one to ensure trustworthiness, the other to assess comprehensibility. Options for quantitative property tests are discussed. Future research should focus on the standardization of testing procedures.

1. Introduction

The development and adoption of complex machine learning (ML) models has given rise to research that seeks to make these models explainable [1, 2, 3]. Explainability methods aim to provide reliable insights into reasons for model behavior. The motivations for why these insights are sought after are manifold: Debugging models during development, finding undesired artifacts in data, learning from models to gain novel scientific insights and to verify that the model fulfills ethical or legal requirements [4]. These insights enable informed decisions about trusting ML models, thus giving a foundation to foster societal acceptance. Despite the multitude of explanation methods available, one open question is still what evaluation process is suitable to quantify performance of an explanation system [5]. For explainable machine learning to evolve as a field we need to find ways to systematically evaluate explanation methods, taking into account both the user and that explanations represent model behavior [6, 7, 8, 9]. In this position paper, we present an evaluation process that factors in both goals and describe directions for quantitative testing to facilitate adoption by practitioners.

*equal contribution

LWDA'22: Lernen, Wissen, Daten, Analysen. October 05–07, 2022, Hildesheim, Germany

katharina.beckh@iais.fraunhofer.de (K. Beckh); semueller@uni-bonn.de (S. Müller)
0000-0002-7824-6647 (K. Beckh); 0000-0002-0778-9695 (S. Müller)

User Independent Properties	
Faithfulness	Refers to how accurate an explanation is to the model's reasoning.
Completeness	Describes how much of the model behavior is described.
Consistency	Refers to how deterministic and implementation-invariant the explanation method is.
Continuity	Similar inputs should have similar explanations.
Contrastivity	Answers "why not?" or "what if?" questions.
Confidence	Probability information in the explanation.
User Dependent Properties	
Compactness	Represents the size of the explanation.
Usability	Degree of fulfillment of user satisfaction and effectiveness.
Plausibility	Refers to how convincing an explanation is to a user.

Table 1
Examples of explanation properties divided into User Independent and User Dependent Properties.

2. Background

Evaluation of explainable systems is typically categorized into three types: (1) application-grounded evaluation with humans and applied tasks (2) human-grounded evaluation with humans but proxy tasks and (3) functionally-grounded evaluation without humans and proxy tasks instead [10].

In a recent review of more than 300 papers from prominent ML conferences, it was found that evaluation is roughly 30% anecdotal evidence and roughly 20% with users [5]. In a smaller-scale survey on explainable natural language processing, 50 papers were reviewed [1]. Roughly 60% used no or "informal" evaluation, 23% comparison to ground truth and 17% human evaluation. The percentage of human evaluation is comparable to the more recent study, however, the number for papers with anecdotal evaluation is double. An underlying reason could be that "no or informal" evaluation includes approaches that were deemed as not anecdotal by the other study. These reviews reveal that at least 30% and up to 60% of papers do not present adequate evaluation for explanation quality. Although explanation methods are abundantly available and incorporated in ML systems, the field is still lacking standards regarding the evaluation of explanation quality similar to the standards we expect from an evaluation on predictive performance.

2.1. Explanation properties

There exist multiple works describing different properties of explanations that can be used for evaluation [7, 11, 12, 13]. Desired properties and definitions are exemplified in the list in Table 1. The list is not exhaustive and what properties (sometimes referred to as attributes or notions [12]) are necessary to consider, depends on the particular application. A frequently discussed distinction is one between evaluating **faithfulness** and **plausibility** of an explanation [7]. Faithfulness is also referred to as correctness or fidelity; plausibility as coherence [5]. Evaluation strategies for these properties differ. For faithfulness, quantitative evaluation metrics exist,

e.g. ”sensitivity” and ”infidelity” [14, 15] for feature attribution methods. Plausibility can be evaluated by comparing ML explanations against ”ground-truth” explanations [16]. For an overview of properties we refer to a recent survey which lists evaluation methods for each property [5]. With a number of identified properties, the question arises why we do not put these properties to use to systematically evaluate explainability methods.

3. Trustworthiness and Comprehensibility

None of the goals of explainable ML, be it generating scientific insight or enabling model verification, can be achieved on the basis of faulty explanations. Hence, it is crucial to ensure that explainability systems deliver a high quality output that represents the model behavior truthfully [6, 7]. At the same time, all explanations are meant to be understood by a human user. Therefore, explanations are only useful if they are comprehensible [8, 9]. An evaluation process for explanation systems will need to ensure both, the quality of the explanations as well as the comprehensibility for the target user. We distinguish between user independent and user dependent properties with respect to the evaluation process (see subsequent sections). For this, we define the notion of **trustworthiness** to encompass all properties that measure the degree to which the explanations reflect the true model behavior. Analogously, we define **comprehensibility** to encompass all properties that measure to what degree and how efficiently the user is able to reach their interaction goals.

3.1. Ensuring Trustworthiness: User Independent Properties (UIPs)

Human preference for an explanation does not necessarily correlate with its trustworthiness [16]. To protect against user bias, we argue that a positive assessment of UIPs needs to be a pre-requisite for the subsequent user-dependent evaluation of the system.

Not all properties are applicable to all explanation methods. For example, explanation methods that involve random sampling are naturally less consistent than deterministic methods. One idea is to use fact sheets to track applicable properties of an explanation method [13], thus making it easier to gauge what properties are relevant for a given system.

We have the impression that existing evaluation measures for UIPs are not widely used yet, but hope that the integration of measures into emerging coding-libraries[1] has an effect on practice.

3.2. Ensuring Comprehensibility: User Dependent Properties (UDPs)

Carrying out user studies is a costly and time-consuming process. We argue that generating reports on UDPs should be made as easy as possible for practitioners. We thus want to discuss possibilities to facilitate quantitative measurements of UDPs.

Compactness is a property that is closely tied to managing users’ *Cognitive Load* [17]. For this, we can turn to the concept of *Working Memory* [18] that describes the limit of how many *chunks* (meaningful items/ thoughts) a human can actively entertain at the same time. What

[1] e.g. Captum and Quantus

constitutes a chunk in the respective application needs context-level definition. For text data it might be a single word, for image data it might be all pixels belonging to the same object. What constitutes a chunk also depends on user experience in the respective domain, e.g. a more experienced user having developed a more complex mental model to associate information with can process larger amounts of information than inexperienced users.

Plausibility measures to what extent the explanation corresponds to user expectations. Providing practitioners with a ground truth dataset containing human verified explanations would enable them to report a degree of explanatory overlap as explored by [16]. Remarkably, the study also found user biases in the subjective rating of explanations. This indicates that purely qualitative evaluation is not sufficient providing another reason to develop quantitative measures. A possible disadvantage of ground truth datasets is that they are task specific, hence, transferability might be an issue. Another direction for plausibility checks are test scenarios, inspired by software engineering practices [19]. Test scenarios can be generated according to user expectations. In a scenario with text data and feature highlighting, possible checks could be the following: "words with low information value, e.g. *the*, should not be part of an explanation" or "replacing adjectives with synonyms should yield the same attribution". However, usefulness of ground truth datasets or other test scenarios have their limits. If an explanation does not comply with user expectations that does not mean that it is not a faithful description of model behavior. Discrepancies between expected explanation and observed explanations have to be investigated carefully as to not blame the explainability method when it is really the model that behaves unexpectedly. If the explanation is trustworthy, a plausibility check does not reveal anything about the explanation but only something about the model.

Usability is the most elusive property of the three because it not only depends on the explanation methods used but also on presentation, UI, tech-affinity of the user and the specific questions the user wants to be answered. Options to quantitatively measure usability are time-on-task or task-success with an explanation condition (with or without explanation) [20, 21, 22]. The challenge remains that user studies, whether they are qualitative or quantitative, require time and resources.

This list of properties is non-exhaustive and relevance of individual metrics may vary depending on the application setting or the diversity of the targeted user group.

4. Process for Human-Grounded Quantitative Evaluation

To establish good evaluation practices we propose the evaluation process shown in Figure 1. The key idea of this process is to enforce UIP evaluation both before evaluating UDPs and also after any explanation changing alterations to the system.

At first the **Context** needs to be captured [23]. **User Interviews** provide an investigation into what goals the users want to accomplish and what questions need to be answered, e.g. [24]. Beyond immediate user intent, engineers may rely on established **Design Guidelines** for comparable scenarios to inform initial design decisions, e.g. regarding compositionality.

This informs the **Requirement Catalogue** used to determine what **Explainability Methods** should be selected to fulfill identified user needs. Upon implementation, the system is evaluated with regards to applicable UIPs. If UIPs indicate an improper representation of model behavior

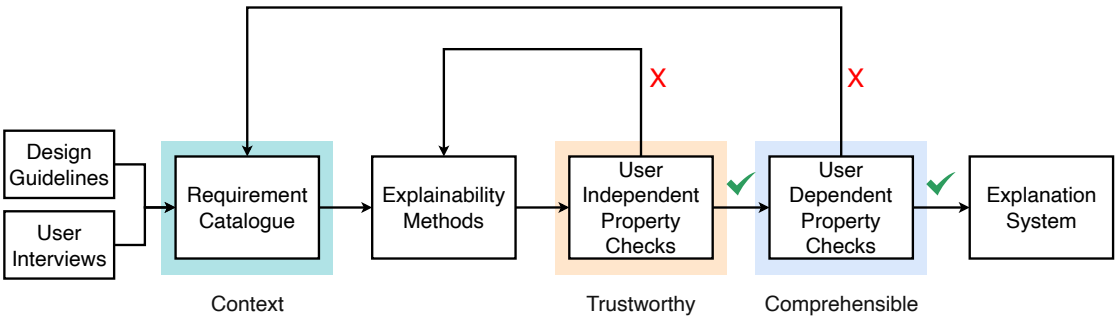

Figure 1: Human-grounded evaluation process for explanation systems.

by the explanations, the engineer returns to choosing different Explainability Methods and UIP checks are repeated. Once UIP checks are passed, the system is assessed with regards to the UDPs. This can involve a real user study or a quantitative evaluation. If the explanations are deemed incomprehensible, this requires an update of the Requirement Catalogue. Thus, UDPs can impact UIPs: For example, if tests failed for Compactness and the user cannot cope with the amount of information presented, displaying less information has an impact on Completeness of the displayed explanation. If the system is adapted, a re-evaluation of the UIPs is necessary to ensure trustworthiness is maintained.

Most likely, UIP-UDP-trade-offs will be unavoidable in practice, e.g. due to specific user needs or a lack of choice of better explanation methods. Finding a balance will be difficult. Returning to our previous example: A UDP check indicates that the user prefers a more compact explanation. The engineer implements a slider to adjust the amount of chunks to be displayed at the same time. Seemingly, all information is still available to the user, yet the user can now pick at what level of compactness the explanation appears the most plausible. This rather enables them to confirm their own bias of the model instead of investigating the true model behavior. Providing the user with information about how their actions affect trustworthiness might be a possible remedy. However, in some scenarios, e.g. where an explanation system is used to investigate a model for certification or in a legal dispute, trustworthiness demands may be nonnegotiable.

In general, we see the risk that compromising UIPs, such as faithfulness, opens the way for manipulation in adversarial settings [25]. We expect that defining the boundaries of permissible UIP-UDP-trade-offs will be challenging as indicated by literature discussing faithfulness-plausibility trade-offs [7, 26].

For works that propose a novel explanation method without a direct use case, it would be helpful to provide information on applicable UIPs and a discussion of how UDPs might impact the performance under which circumstances.

”The food was delicious and the price somewhat fair.”	If a user argues for more compactness the explanation could look like the following sentence with ”fair” not being highlighted.
”The food was delicious and the price somewhat fair.”	By removing the term ”fair” from the feature highlighting the explanation is more compact but less complete.

Example: Interaction between the UDP Compactness and the UIP Completeness.

5. Open Research Directions

The standardization of UIPs and UDPs would enable the development of software-packages that automatically test for applicable properties given an explanation system.

The number of quantitative tests available for UDPs is still low which we attribute to the subjective nature of UDPs. Nevertheless, quantitative tests for UDPs will be helpful and reliable means to better compare methods.

Furthermore, we recommend that newly proposed explanation methods perform ablation studies to simulate possible application scenarios, or at least include discussions to analyze trustworthiness-comprehensibility trade-offs. Likewise, indicating applicable UIPs and reporting on respective statistics should become the standard, similar to the reporting that is customary for predictive performance.

Trust is often stated as goal of explainable ML. Naturally, it would be interesting to investigate how explanation systems developed with the proposed process score with regards to (in-) appropriate trust.

6. Conclusion

Evaluation of explainable ML is essential to demonstrate trustworthiness and ensure comprehensibility of an explanation system. However, there are no guidelines or standards for evaluation yet. In this work, we argued for quantitative human-grounded evaluation of explainable ML. We presented an evaluation process that includes the user and places importance on quantitative tests. Explanation quality is considered as a construct consisting of several properties. We split the properties into two sets, one to ensure trustworthiness, the other to assess comprehensibility. We highlight that adaptations of the explanation system based on comprehensibility checks can impact properties that quantify trustworthiness. We raise the question whether this trade-off is justifiable under any circumstances and call for a discussion within the community to further develop a position.

Acknowledgments

We thank the ML2R Trustworthy ML group and the MLAI Lab for the helpful discussions. This research has been partly funded by the Federal Ministry of Education and Research of Germany as part of the competence center for machine learning ML2R (01|S18038BC). K. Beckh

has contributed as part of the SmartHospital.NRW project which is funded by the Ministry for Economic Affairs, Innovation, Digitalization and Energy of the State of North Rhine-Westphalia.

References

[1] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, ACL, Suzhou, China, 2020, pp. 447–459.

[2] K. Beckh, S. Müller, M. Jakobs, V. Toborek, H. Tan, R. Fischer, P. Welke, S. Houben, L. von Rueden, Explainable machine learning with prior knowledge: An overview, arXiv preprint arXiv:2105.10172 (2021).

[3] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[4] C. Molnar, Interpretable machine learning, second ed., leanpub.com, 2022.

[5] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, arXiv preprint arXiv:2201.08164 (2022).

[6] D. Alvarez-Melis, T. S. Jaakkola, On the robustness of interpretability methods, arXiv preprint arXiv:1806.08049 (2018).

[7] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 2020, pp. 4198–4205.

[8] M. O. Riedl, Human-centered artificial intelligence and machine learning, Human Behavior and Emerging Technologies 1 (2019) 33–36.

[9] A. Páez, The pragmatic turn in explainable artificial intelligence (xai), Minds and Machines 29 (2019) 441–459.

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

[11] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[12] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Information Fusion 76 (2021) 89–106.

[13] K. Sokol, P. Flach, Explainability fact sheets: A framework for systematic assessment of explainable approaches, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, ACM, 2020, p. 56–67.

[14] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for deep neural networks, in: International Conference on Learning Representations, 2018.

[15] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, P. K. Ravikumar, On the (in)fidelity and sensitivity of explanations, in: Advances in Neural Information Processing Systems, volume 32, 2019.

[16] S. Mohseni, J. E. Block, E. Ragan, Quantitative evaluation of machine learning explanations: A human-grounded benchmark, in: 26th International Conference on Intelligent User Interfaces, 2021, pp. 22–31.
[17] J. L. Plass, R. Moreno, R. Brünken, Cognitive load theory, Cambridge University Press (2010).
[18] N. Cowan, The magical mystery four: How is working memory capacity limited, and why?, Current Directions in Psychological Science 19 (2010) 51–57.
[19] M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh, Beyond accuracy: Behavioral testing of NLP models with CheckList, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 2020, pp. 4902–4912.
[20] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, F. Doshi-Velez, Human evaluation of models built for interpretability, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 7, 2019, pp. 59–67.
[21] J. P. Dietvorst, Berkeley J .and Simmons, C. Massey, Algorithm aversion: People erroneously avoid algorithms after seeing them err, Journal of Experimental Psychology: General 144 (2015) 114.
[22] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, H. Wallach, Manipulating and measuring model interpretability, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, ACM, 2021.
[23] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, J. D. Weisz, Expanding explainability: Towards social transparency in ai systems, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–19.
[24] Q. V. Liao, D. Gruen, S. Miller, Questioning the AI: Informing Design Practices for Explainable AI User Experiences, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (2020) 1–15. URL: http://arxiv.org/abs/2001.02478. arXiv:2001.02478.
[25] S. Bordt, M. Finck, E. Raidl, U. von Luxburg, Post-hoc explanations fail to achieve their purpose in adversarial contexts, arXiv preprint arXiv:2201.10295 (2022).
[26] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., Queue 16 (2018) 31–57.