XAI 2.0: The multidisciplinary conundrum of eXplainable Artificial Intelligence, its interdisciplinarity and a transdisciplinary path forward

Riccardo Guidotti
University of Pisa, Italy
riccardo.guidotti@unipi.it

1 Evaluating Explanation Methods

1.1 Problem Description

In the last years, we are observing an exploding number of proposals for eXplainable Artificial Intelligence (XAI), either post-hoc explanation methods but also interpretable-by-design Machine Learning (ML) approaches [8, 3, 1, 4, 2]. However, despite the broad interest in the design of XAI methods and the proven effectiveness of some of them, there is still an enormous lack of standardizing how much an explanation method is good and which are the objective minimum requirements for explainability. To make a parallelism with traditional classification and regression tasks ML, what is needed for XAI is an objective evaluation measure like the accuracy or the mean squared error.

1.2 Challenges and Limitations

Evaluating explanation methods in XAI is a complex task mainly due to the lack of a shared and universally accepted definition of explanation. As a consequence, it difficult to compare different types of explanations and different types of explanation methods to determine which are the most effective.

A limitation of current evaluations of XAI methods is that they typically analyze peculiarities of the explanation methods without accounting for the interaction with the final user [3, 6, 7]. This is obviously a starting point but this kind of evaluation completely misses to judge the impact with the human that is willing to understand the reasons of the decision.

On the other hand, evaluation of the XAI methods involving humans is typically prone to bias and errors because the number of participants in the human evaluation study is generally not representative of the population and thus, the results may not generalize to other situations. In [9], a survey of user studies, is shown that only 36 out of 127 papers analyzed regarding counterfactual explainers conducted any form of human evaluation, and only 7% of them competitively test alternative algorithms. Furthermore, these studies are corrupted by poorly-reproducible experiments and inappropriate statistical analyses. Hence, there are no studies reporting solid evidence on the efficacy of XAI methods w.r.t. human participants. In addition, users are typically "passive recipients" of the explanations, and it is not tested if they can use/exploit them, basing the evaluation only on a subjective or passive judgment. Indeed, most studies try to understand whether or not an explanations impact the user behavior by testing if the supply of an explanation to an automatic AI decision has any effect on user behavior by comparison to no-explanation controls [5, 10, 11]. Very few studies compare in such a way more than one explanation method.

Another challenge is that the quality of an explanation is often subjective, as it depends on the perspective of the person interpreting the explanation. For example, a human expert in a particular field may have different criteria for evaluating an explanation than a layperson. Additionally, different individuals may have different levels of prior knowledge and understanding of a given subject, which can also influence their perception of the quality of an explanation. Furthermore, the complexity of the task that the XAI model is performing can also affect the evaluation of the explanation. For example, an explainer that is providing explanations for a simple task may be evaluated differently than an explainer that is providing explanations for a more complex task.

1.3 Future Solutions

From the analysis of the previous section, we can understand that the XAI community might not be taking the right approach to judge explanations and explanation methods. An intuition might be that the strategies adopted in the XAI assessment are quite *method-centered* or *explanation-centered* and not sufficiently *human-centered* or *task-centered*. Thus, adopting such perspectives may shed a different light on the current knowledge of how people understand explanations.

Thus, it might be better to judge to which extent the explanation returned by an XAI method for a certain task is actually the one that the users really require/expect or not. A possibility to do it, that is also in line with the path started by some researchers w.r.t. synthetic objective evaluations [7], is to gather users-generated ground-truth of explanations to be used both for designing human-centered explanations, ML-inspired explainers, and more importantly, to effectively and objectively judge existing explainers.

References

[1] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[2] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115, 2020.

[3] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo. Benchmarking and survey of explanation methods for black box models. *CoRR*, abs/2102.13076, 2021.

[4] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[5] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In *IUI*, pages 275–285. ACM, 2019.

[6] R. Guidotti. Evaluating local explanation methods on ground truth. *Artif. Intell.*, 291:103428, 2021.

[7] R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

[8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019.

[9] M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. *arXiv preprint arXiv:2103.01035*, 2021.

[10] A. Lucic, H. Haned, and M. de Rijke. Why does my model fail?: contrastive local explanations for retail forecasting. In *FAT**, pages 90–98. ACM, 2020.

[11] C. Metta, R. Guidotti, Y. Yin, P. Gallinari, and S. Rinzivillo. Exemplars and counterexemplars explanations for skin lesion classifiers. In *HHAI*, volume 354 of *Frontiers in Artificial Intelligence and Applications*, pages 258–260. IOS Press, 2022.