Explaining is not enough: Four open problems in Explainable AI for AI with user interaction

Andreas Theissler

Aalen University of Applied Sciences, Aalen, Germany
andreas.theissler@hs-aalen.de

Abstract

Artificial intelligence (AI), specifically machine learning, has found its way into everyday life making it more important than ever to find ways to assess the used machine learning models. Due to their currently observed superior performance, black-box models are used in many cases. In the assessment of these, interpretation of the models or their decisions is one key element. Hence, in this extended abstract we eloborate on the research field of explainable AI (XAI) identifying four open problems that are specifically relevant for AI systems with user interaction. We formally describe and motivate the problems and propose to work on these to contribute to the overriding goal of trustworthy AI.

1 Explainable AI: A brief introduction

With the increased performance of machine learning (ML) models, their complexity has also increased. Many of the currently successful ML models are "black box models", hence, are not interpretable. The lack of **explainability** or **interpretability** may prevent end users from trusting the models, preventing their use in industrial or medical settings (Langone, Cuzzocrea, and Skantzos 2020; Dwivedi et al. 2023; Raab, Theissler, and Spiliopoulou 2022). As a consequence, the research field of explainable AI (XAI) has emerged in recent years. (Biran and Cotton 2017) state "systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation." Hence, a distinction can be made between ML models providing interpretability, i.e. **intrinsically interpretable models**, and those requiring explanations, i.e. **post-hoc explanations**, to become interpretable.

According to survey and position papers (Adadi and Berrada 2018; Guidotti et al. 2018; Theissler et al. 2022), post-hoc explanations can be further categorized into **global methods** explaining the entire ML model and **local methods** explaining model predictions for individual instances. Moreover, XAI methods (**explainers**) can be **model-agnostic** or **model-specific**: Model-agnostic methods can be applied to any ML model, given that inputs and outputs match the XAI method (Barredo Arrieta et al. 2020). In contrast, model-specific methods access the ML models' internals and are thereby restricted to specific types or classes of ML models.

2 Our notation

We denote an ML model as M_j and its prediction for a single data instance x as $M_j(x)$. Based on that we wish to establish the following notation: We denote a post-hoc explanation as ϵ_i. More specifically we denote a local explanation, i.e. for a single data instance, as $\epsilon_i(M_j(x))$ and a global explanation $\epsilon_i(M_j)$ (see Table 1).

Furthermore we introduce $u(\epsilon(...))$ which refers to the – hard to quantify – function of users *understanding an explanation*, i.e. $u(\epsilon_i(M_j(x)))$ refers to understanding a local and $u(\epsilon_i(M_j))$ a global explanation.

3 Four open problems

Four open problems in the field of XAI that are specifically relevant for AI systems with user interaction are identified.

3.1 The non-inherent semantics problem

For a data instance x, local explanations uncover which parts, e.g. pixels in images or time points in time series, influenced a prediction $M_j(x)$. However, we argue that these types of explanations – for example attribution methods – show *where* the relevant parts in the input data are, but *not why* the prediction was made. Following our notation, we can state that

$$\epsilon_i(...) \neq u(\epsilon_i(...)) \tag{1}$$

There are applications for which the localization of influential parts in the data can act as an explanation, e.g. for standard images or in some cases for text, justifying the use of these well-developed explainers. Yet, there is a wide range of applications where the pure localization is not a sufficient explanation, e.g. for the majority of time series data (Theissler et al. 2022) and for non-self-explanatory images. We term this as the **non-inherent semantics problem**.

3.2 The variance problem

For a set of *well-trained* ML models M_i we generally expect the models to yield comparable performance on a representative data set X. Thus, it seems reasonable to demand the same for explanations ϵ_i explaining these models. However, we observe that we get rather different explanations ϵ_i from different explainers. A recent experimental study (Bodria

AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems. Stanford University, Stanford, CA, USA.

	model-agnostic	**model-specific**
local	$\epsilon_i(M_j(x))) \quad \forall j \quad \text{where} \quad i = const$	$\epsilon_i(M_j(x))) \quad \forall i,j \quad \text{where} \quad i = j$
global	$\epsilon_i(M_j)) \quad \forall j \quad \text{where} \quad i = const$	$\epsilon_i(M_j)) \quad \forall i,j \quad \text{where} \quad i = j$

Table 1: Proposed notation allowing a compact description for the common categorization of post-hoc explanations.

et al. 2023) showed that XAI methods may yield highly variant explanations for the same underlying ML model, suggesting different parts of the input data to be the potential cause for a given prediction. So, we can state that for currently used explainers

$$\epsilon_i(\dots) \overset{!}{\approx} \epsilon_j(\dots) \quad \forall i,j \tag{2}$$

We refer to this as the **variance problem**.

3.3 The dynamic target user problem

The need to tailor explanations ϵ_i with the target users in mind (e.g. ML engineers, domain experts, or laypersons) is commonly agreed on.

However, from our point of view, there is one aspect to be added: The capabilities of users to understand the explanations, i.e. $u(\epsilon(...))$, are to be seen dynamic: (a) users have learning curves and (b) users understand in hierarchical ways, i.e. there is need for *adaptive explanations* which – for the same user – should also adapt w.r.t. experience (or time) t. Hence, we could state

$$u(\epsilon_i(...), t) \tag{3}$$

We refer to this as the **dynamic target user problem**.

3.4 The evaluation problem

We like to point out that the generation of an explanation ϵ_i is not sufficient if ϵ_i is not understood by the targer users. While this fact is well-known, we like to postulate that this should be addressed when using or proposing XAI methods. Based on our notation, one may demand to evaluate

$$u(\epsilon_i(M_j(x))) \quad \text{or} \quad u(\epsilon_i(M_j)) \tag{4}$$

In (Doshi-Velez and Kim 2017), evaluation is subdivided into threee levels: functional grounded (e.g. quantitative metrics) and two levels of evaluations involving users (laypersons and experts). However, evaluations involving users are often avoided in papers, e.g. out of approx. 60 XAI methods reviewed in (Theissler et al. 2022) four were evaluated with a user study. We observe that the plausibility of explanations is often shown with a small set of examples (including work by ourselves).

Research on quantitative metrics has made progress as shown in (Hedström et al. 2023; Schlegel and Keim 2023), but $u(\epsilon_i(...))$ is hard to quantify. We believe that for the time being users should be involved in the evaluation alongside the use of quantitative metrics. If substantial progress is made in studying the human interpretability of XAI explanations quantitatively, they *might* be able to replace user tests in the future *for recurring cases*. We refer to this as the **evaluation problem**.

References

Adadi, A.; and Berrada, M. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6: 52138–52160.

Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

Biran, O.; and Cotton, C. V. 2017. Explanation and Justification in Machine Learning : A Survey. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*.

Bodria, F.; Giannotti, F.; Guidotti, R.; Naretto, F.; Pedreschi, D.; and Rinzivillo, S. 2023. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 1–60.

Doshi-Velez, F.; and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*.

Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; and Ranjan, R. 2023. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.*, 55(9).

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5): 1–42.

Hedström, A.; Weber, L.; Krakowczyk, D.; Bareeva, D.; Motzkus, F.; Samek, W.; Lapuschkin, S.; and Höhne, M. M.-C. 2023. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34): 1–11.

Langone, R.; Cuzzocrea, A.; and Skantzos, N. 2020. Interpretable Anomaly Prediction: Predicting anomalous behavior in industry 4.0 settings via regularized logistic regression tools. *Data & Knowledge Engineering*, 130: 101850.

Raab, D.; Theissler, A.; and Spiliopoulou, M. 2022. XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series. *Neural Computing and Applications*, 1–18.

Schlegel, U.; and Keim, D. A. 2023. A Deep Dive into Perturbations as Evaluation Technique for Time Series XAI. In *World Conference on Explainable Artificial Intelligence*, 165–180. Springer.

Theissler, A.; Spinnato, F.; Schlegel, U.; and Guidotti, R. 2022. Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions. *IEEE Access*, 10: 100700–100724.