# Position: XAI needs formal notions of explanation correctness

**Stefan Haufe**
Technische Universität Berlin
Physikalisch-Technische Bundesanstalt, Berlin
Charité – Universitätsmedizin Berlin
haufe@tu-berlin.de

**Rick Wilming, Benedict Clark, Rustam Zhumagambetov, Danny Panknin, Ahcène Boubekki**
Physikalisch-Technische Bundesanstalt, Berlin

## Abstract

The use of machine learning (ML) in critical domains such as medicine poses risks and requires regulation. One requirement is that decisions of ML systems in high-risk applications should be human-understandable. The field of "explainable artificial intelligence" (XAI) seemingly addresses this need. However, in its current form, XAI is unfit to provide quality control for ML; it itself needs scrutiny. Popular XAI methods cannot reliably answer important questions about ML models, their training data, or a given test input. We recapitulate results demonstrating that popular feature attribution and counterfactual estimation methods systematically attribute importance to input features that are independent of the prediction target, and that popular faithfulness metrics incentivize attribution to such features. This limits their utility for purposes such as model and data (in)validation, model improvement, and scientific discovery. We argue that the fundamental reason for this limitation is that current XAI methods do not address well-defined problems and are not evaluated against objective criteria of explanation correctness. Researchers should formally define the problems they intend to solve first and then design methods accordingly. This will lead to notions of explanation correctness that can be theoretically verified and objective metrics of explanation performance that can be assessed using ground-truth data.

## 1 Introduction

The use of machine learning (ML) holds great promise in many fields including high-risk domains such as medicine. Regulations like the European AI Act demand that "high-risk AI systems shall be designed and developed [...] to enable deployers to interpret the system's output and use it appropriately." (European Commission, 2021). This need for "human-understandable" descriptions of the functions implemented by individual ML models is seemingly addressed by the field of "explainable artificial intelligence" (XAI). However, the formal basis of XAI is underdeveloped. Consequently, the possibility of using XAI for ML quality assurance is currently strongly limited. While serious conceptual and technical shortcoming of existing XAI methods and paradigms have been voiced by numerous authors (see Discussion section), we here focus on the fitness of feature attribution and counterfactual estimation methods for various specific purposes purported in the literature. Using two minimal counter-examples involving so-called suppressor variables, we show that popular methods cannot systematically answer questions relevant to serving these purposes.

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

## 2 Purported purposes of XAI

The popularity of XAI tools rests on their promise to provide insight into the properties of ML models, their training data, a given test input submitted to the model, and/or the interplay between these. A particularly popular class of XAI methods are those that attribute an "importance" score to each feature of a given test input. It has been argued that such feature attribution methods can be used for:

**Model (in)validation:** It is often of interest to know which features of a dataset or single sample an AI system "bases" its decision on. This would then be used to "validate" a model. In mammographic data analysis, a radiologist would likely trust a cancer diagnosis made by an AI if told that the decision was based on a patch of tissue they themselves identify as cancerous. Conversely, if the XAI method assigns high "importance" to features that are known not to be associated with cancer, this might lead to the dismissal of the model as being wrong (Saporta et al., 2022). Ribeiro et al. (2016) state "A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the 'flu' prediction, while 'no fatigue' is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.".

**Data (in)validation:** Similarly, one may be interested in whether a model bases its decisions on confounding variables. Confounders induce correlations between training in- and outputs that can be used by the model for prediction. However, the same correlations may not be present in a testing context, leading the model to perform poorly. Lapuschkin et al. (2019) use XAI methods in the process of identifying a watermark in images that acts as a confounder. The intended purpose of XAI, in this case, is, thus, to perform quality control not only for models but also data.

Beyond diagnostic purposes, XAI has also been suggested to enable actionable consequences such as:

**Scientific discovery:** Various authors (Samek and Müller, 2019; Jiménez-Luna et al., 2020; Tideman et al., 2021; Watson, 2022; Wong et al., 2024) argue that XAI methods could be used to discover novel associations between variables, generating new hypotheses that could be tested in future experiments. For example, a disease might be related to a complex interaction of multiple previously unknown genetic factors. Such an interaction might be implicitly used by an ML model. The promise of XAI methods is then to identify the features contributing to the interaction.

**Identification of intervention targets:** It is frequently assumed that XAI could be used to identify features, the manipulation of which would change a model's output, a task also known as algorithmic recourse (Ustun et al., 2019). For example (see Ustun et al., 2019), a bank might use an ML model to predict the return probability of a loan. For a known model and a given input, XAI would then be able to recommend changes of input variables (e.g., 'increase salary') to turn a negative outcome into a positive one. Similarly, it is assumed that XAI can help to verify that protected attributes (e.g., gender, race) do not influence model decisions. In an intensive care unit, an ML model might be used to predict mortality or other severe outcomes. Using XAI to identify possible intervention targets, such as medications, in this context (e.g., Ates et al., 2021) goes beyond algorithmic recourse as the underlying assumptions is that interventions have real-world consequences on the target variable beyond just changing the model output.

## 3 Existing feature attribution methods do not serve important purposes

It is indisputable that the purposes of XAI presented in Section 2 are of high relevance. For these, it is assumed, explicitly or implicitly, that XAI methods have the following capabilities.

**A1**: Global feature attribution methods (characterizing a model independent of a specific input) identify data features that are statistically associated with or even causally related to the model output or target variable (see, e.g., Ribeiro's example cited above, Ribeiro et al., 2016).

**A2**: Local feature attribution methods (characterizing a model's application to a specific input) identify features the manipulation of which is, a), feasible and would influence, b), the model output or, c), even the prediction target in the real world.

Research has shown, though, that the assertions listed above do not hold in general for several of the most popular XAI methods. Thus, these methods fail to reliably serve the purposes mentioned in Section 2. We will demonstrate this next using two minimal examples.

### 3.1 Two minimal examples of classification problems

In the following, non-scalar values are highlighted in bold and we denote random variables by upper-case symbols, e.g. Z, whereas lower-case analogs, e.g. z, represent their respective realizations.

**Example A:** In Haufe et al. (2014), the two-dimensional classification problem $\mathbf{X} = \mathbf{a}Z + \mathbf{H}, \quad Y = Z$ is introduced, with $\mathbf{a} = (1, 0)^\top$, $Z \sim$ Rademacher$(1/2)$, and $\mathbf{H} \sim N(\mathbf{0}, \mathbf{\Sigma})$ with covariance $\mathbf{\Sigma} = \begin{pmatrix} s_1^2 & cs_1s_2 \\ cs_1s_2 & s_2^2 , \end{pmatrix}$, where $s_1$ and $s_2$ are non-negative standard deviations, and $c \in [-1, 1]$ is a correlation. In this example, only feature $X_1$ is correlated with the classification target $Y = Z$ through $a_1 = 1$. In contrast, $X_2$ is independent of $Y$ since $a_2 = 0$. Both features are correlated through the superposition of additive noise $\mathbf{H}$ with covariance $\mathbf{\Sigma}$. A depiction of data generated under the model is provided in Figure 1 (a/b) in Appendix A. For $c \neq 0$, the Bayes-optimal bivariate linear classification model $f_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ can reduce the contribution of $\mathbf{H}$ from $X_1$ using information contained in $X_2$, and thereby estimate $y$ as $\hat{y} = f_{\mathbf{w},b}(\mathbf{x})$ more precisely as what would be possible using $X_1$ alone (Haufe et al., 2014). To this end, it needs to put non-zero weight $w_2 = -\alpha c s_1/s_2$ on $X_2$, where $\alpha = (1 + (cs_1/s_2)^2)^{-\frac{1}{2}}$ and $||\mathbf{w}||_2 = 1$. This shows that linear models can assign arbitrarily high weights on features, like $X_2$, that have no statistical association with $Y$.

**Example B:** An even simpler example is given by the generative model $X_1 = Y - X_2$, where the suppressor $X_2$ and the target $Y$ are independent (Haufe et al., 2014). Here the Bayes-optimal linear model with weights $w_1 = w_2 = 1$ completely removes the nuisance term $X_2$ from $X_1$ to recover $Y$, yielding a model output that is statistically independent of $X_2$. Such examples challenge the concept of a model "using a feature" or "basing its decision on a feature".

### 3.2 Suppressor variables

Features like $X_2$ in Examples A and B, which improve predictions without being predictive themselves, are called suppressor variables in causal terminology (Conger, 1974). Causal diagrams of the generative models in both examples are provided in Figure 1 (c) in Appendix A. Broadly speaking, any variable that is not informative (statistically associated with the target) itself but statistically related with an informative variable (e.g. modulating it through an independent mechanism) is a suppressor. Suppressors occur widely in real-world datasets and hamper model interpretations. As one example, the prevalence of a disease may be related to a person's blood pressure but not their age. However, as blood pressure has an age-dependent baseline, the model might need to adjust its prediction with respect to that baseline in order to remove irrelevant variance introduced by age. Age, thereby, become a suppressor variable. In image classification, non-discriminative features such as lighting or weather conditions, or non-discriminative objects occluding class-specific objects, could be suppressors. It is, thus, possible for suppressors to act in highly non-linear ways and it is also possible to find examples of suppressors at higher semantic (concept) levels.

### 3.3 Existing feature attribution methods attribute importance to variables unrelated to target

Recent theoretical and empirical research has shown that various popular feature attribution methods consistently assign importance to suppressor variables (Haufe et al., 2014; Kindermans et al., 2018; Wilming et al., 2022, 2023; Clark et al., 2024). We call these methods suppressor attributors. Kindermans et al. (2018) showed analytically that the importance scores returned by gradient-based techniques (Baehrens et al., 2010), LRP (Bach et al., 2015), and DTD (Montavon et al., 2017) reduce to the weight vector $\mathbf{w}$ in case of linear models. Thus, these methods are suppressor attributors. In Wilming et al. (2023), the latter was shown also for Shapley values (Shapley, 1953) and their approximations such as SHAP (Lundberg and Lee, 2017; Aas et al., 2021), as well as for LIME (Ribeiro et al., 2016), integrated gradients (Sundararajan et al., 2017), and counterfactual explanations (Wachter et al., 2017). A list of suppressor attributing methods is provided in Table 1 in Appendix A.

### 3.4 Existing feature attribution methods violate common assertions

Since suppressor variables have no statistical or causal association with the target variable, suppressor attributors violate assertion 3, which has implications regarding their expected utility for the purposes introduced in Section 2. Suppressor features may often not coincide with prior expectations of an expert. Therefore, suppressor attributors cannot be used in a straightforward way to validate models or models' decisions using expert knowledge as insinuated by Ribeiro et al. (2016). Moreover, since it cannot be concluded that the highlighted features are part of previously unknown interactions or are causally related to the output, these methods cannot be reliably used to facilitate scientific discoveries or to invalidate models. For example, high importance on a protected attribute does not necessarily mean that the method "uses" this attribute for prediction. The model may also just remove variance related to that attribute from other, informative, variables. Finally, a prerequisite for identifying confounding variables causally influencing both in- and outputs of a model is to be able to recognize features with a statistical association to the target in the first place. The inability of suppressor attributors to distinguish such features from suppressor variables, as discussed here, prevents XAI methods from answering causal questions of confounding.

Another critical observation from Example B is that $X_2$ cannot be independently manipulated but only jointly with $X_1$, due to its causal influence on $X_1$. Similarly, $X_1$ and $X_2$ are coupled through the common noise $\mathbf{H}$ in Example A. Depending on the nature of $X_2$, it then might not be possible to intervene at $X_2$ alone by setting its value independent of $H$. In both examples, any causally feasible intervention (through $X_2$ or $H$) would have no effect on $Y$ in the real world. In Example B, it would not even affect the model output, as the model is invariant to changes in $X_2$ by construction. In Example B, possible interventions on $X_2$ could affect the model output; however, not in ways that would correlate with changes in $Y$. Such interventions would create out-of-distribution data that do not follow the data generating process (similar to pertubations based faithfulness metrics discussed below). A discussion of the effect of such pertubations on the model output, therefore, does not appear to be useful with respect to understanding the model.

Both examples also illustrate that suppressor attributors also violate assertion 3 with the consequence that they cannot reliably suggest feasible intervention targets. The insufficiency of counterfactual explanations to enable algorithmic recourse has been laid out in detail in Karimi et al. (2021).

## 4 Structural limitations of current XAI research

The results presented above have been established through joint theoretical analyses of data-generating processes, ML models, and feature attribution methods as well as through simulations using synthetic data with known ground-truth explanations. These techniques are not currently part of the standard toolkit for assessing the quality of explanations and XAI methods, pointing to the following fundamental structural limitations of the field.

**Lack of formal problem definitions** The current XAI terminology uses the term "explanation" indiscriminately in different contexts. This lack of differentiation gives rise to equivocality of evaluation framework and is reflective of a deeper absence of well-defined problems for XAI to solve. Even though XAI methods are frequently proposed to serve purposes such as those listed in Section 2, it is rarely stated what concrete types of conclusions can be drawn from the explanations provided by any particular method, and under which assumptions each conclusion is valid. Instead, various popular XAI methods are purely algorithmically defined without reference to a formal problem or a cost function to be minimized, leading to a logical loop where the method defines the problem it solves. In their work, Ribeiro et al. (2016) do not define what the correct features for LIME to highlight would be – the algorithm itself is considered to be the definition of feature importance.

**Existing theory spares out notions of explanation correctness** Existing theoretical work has postulated axioms that are desirable for XAI methods to fulfill. For example, according to Sundararajan et al. (2017), a method satisfies sensitivity, if a) for every input and baseline that differ in one feature but have different predictions, the differing feature is given non-zero importance, and if also b) the importance of a variable is always zero if the function implemented by the deep network does not depend (mathematically) on it. Axioms like this encode meaningful sanity checks but do not provide a notion of correctness or utility-for-purpose of an explanation.

**XAI methods are mostly ignorant of the data distribution and causal structure** With few exceptions, XAI methods are applied post-hoc to model weights or outputs only. However, a model's behavior cannot be meaningfully interpreted without access to the correlation or causal structure of its training data (Haufe et al., 2014; Weichwald et al., 2015; Karimi et al., 2021; Wilming et al., 2023).

The same model weights that cancel out a suppressor in Examples A and B (see Section 3.1) would have a completely different function, and hence interpretation, when applied to independent features.

Most XAI methods explicit or implicit assume independent features. This reflects the common conception that the sole mechanism by which multivariate models achieve their predictive power is to combine (independent) information in order to leverage non-linear interactions in the data. This perspective overlooks that an equally important task of multivariate models is to denoise interrelated features, which is achieved by *removing* task-irrelevant signals. Incorrectly assuming feature independence can lead to violations of assertion 3 and assertion 3, and, thereby, to all of the described misinterpretations. On the other hand, the same interpretations can often be shown to be valid if features are indeed independent. Karimi et al. (2021), for example, show that correct algorithmic recourse is possible under feature independence. Various methods do take the dependency structure of the training data into account but typically make further limiting assumptions such as Gaussianity or linearity. Pattern (Haufe et al., 2014) and PatternNet/PatternAttribution (Kindermans et al., 2018) can be shown to correctly reject suppressor variables in the studied Examples A and B, in which these assumptions hold. But these methods are currently either inapplicable to non-linear models or show sub-par performance in benchmarks involving non-linear data (Clark et al., 2024).

**"Interpretable" models share limitations of XAI** Various authors (e.g., Rudin, 2019) make a distinction between "explainable AI", which would include post-hoc feature attribution methods, and "interpretable AI", which would include model architectures that are "understandable" to humans due to their simplicity. These are also referred to as "glassbox" models (Rai, 2020), and examples include linear models, models with sparse coefficients, and decision trees. However, what exact interpretations such models presumably afford is rarely stated. In the above examples A and B, Bayes-optimal linear models are uniquely defined and assign non-zero weights to suppressor variables, prohibiting certain desired interpretations and precluding certain actionable consequences that would be valid if the features were statistically independent. Thus, simple models can share unrealistic assumptions such as the assumption on independent features with many feature attribution methods.

Given these challenges of correctly interpreting even simple models, an often assumed "tradeoff" between predictiveness and "interpretability" of models (Shmueli, 2010; Del Giudice, 2024) appears questionable. Rather, even simple models cannot be unambiguously interpreted without knowledge of their underlying training data distribution. This is not to say, though, that simple models cannot easen certain interpretations. For example, sparse models can significantly reduce the number of features the behavior of which needs to be investigated. Notwithstanding, sparsity alone does not guarantee that a feature or neuron with non-zero weight is not a suppressor (Haufe et al., 2014).

**Empirical evaluation frameworks also spare out explanation correctness** Existing frameworks for empirical XAI evaluation (e.g., Hedström et al., 2023) often focus on secondary desiderata such as robustness of explanations. Among these, "faithfulness" metrics are often used as surrogates for explanation correctness. Definitions of faithfulness differ and are often not formulated in mathematically stringent form (see, Guidotti et al., 2019; Jacovi and Goldberg, 2020). Practically, the most common operationalization of faithfulness is that the ablation (e.g., omission or obfuscation) of an important feature will lead to a drop in a model's prediction performance. The presence of such a drop is then used to assess "correctness". Popular perturbation approaches include permutation feature importance (Breiman, 2001), stability selection (Meinshausen and Bühlmann, 2010), pixel flipping (Samek et al., 2017), RemOve And Retrain (ROAR, Hooker et al., 2019), and Remove and Debias (ROAD, Rong et al., 2022), and prediction difference analysis (e.g., Blücher et al., 2022). A variation is the model parameter randomisation test (MPRT, Adebayo et al., 2018).

Despite the simplicity and intuitive appeal of faithfulness metrics, Wilming et al. (2023) show that removal or manipulation of $X_2$ in Examples A and B leads to an inevitable decrease in classification performance, which would lead XAI methods attributing high importance to $X_2$ to appear as faithful. This is because faithfulness metrics do not take the data-generating process and the resulting dependency structure in the data fully into account. In that respect, XAI methods and the metrics used to assess their performance share identical limitations.

**Insufficiency of real data and human judgment to validate XAI** Real datasets are often used for empirical evaluations of XAI methods. In such studies, no ground-truth for the inherently unsupervised XAI problem is available. Several studies (e.g., Holzinger et al., 2019; Biessmann and Refiano, 2021) consider human judgments on real data as a surrogate ground-truth for XAI validation, where human experts either annotate inputs ex ante to provide ground-truth explanations or are asked to judge the quality of explanations ex post. While important, such approaches are insufficient as (sole) validations due to the possibility of both Type-I and Type-II errors in human judgments. For example, there may be complex statistical patterns in the data that are leveraged by ML models but (currently) unknown to humans. Similarly, human experts may hold wrong beliefs based on weak prior evidence. Human-computer interaction studies are considered an objective way to quantify the added value of AI explanations (e.g., Jesus et al., 2021). Such studies compare the joint performance of a human user with access to an XAI with the performance of the user knowing only the outcome of the AI's prediction, the performance of the user alone, and the performance of the AI alone. However, even a practical advantage of XAI does not automatically predict XAI correctness, as there is a possibility of circular reinforcement of wrong beliefs, whereby the human may adapts their judgment to previously provided incorrect explanations. In very specific cases, this could be avoided by providing explanations that can be verified by the user (Fok and Weld, 2023).

**Algorithm-driven development** A common paradigm of XAI development is to start with the design of an algorithm and then to try to prove its utility for various purposes by applying it to selected datasets and models. Unfortunately, this approach opens the door to biases due to implicit subjectivity in the choice of the experiments performed and reported. Thereby it becomes possible that properties inferred by XAI methods are spurious or trivial. For example, the finding that an image feature highlighted by XAI is also a confounder could be an anecdotal coincidence rather than reflecting a systematic capability of the XAI method to identify confounders.

## 5 Towards XAI for quality assurance

Apart from the points raised here, XAI methods have been criticized in many further ways (e.g., Ghassemi et al., 2021; Sokol and Flach, 2020; Weber et al., 2024). For example, the low robustness of XAI explanations has been noted (Babic et al., 2021). Explanations provided by different XAI methods are often found to be inconsistent. This can be used by an adversary (e.g., the provider of an ML algorithm in need to explain a decision to a user) to provide arbitrary explanations (Bordt et al., 2022). Similarly, developers of XAI methods could present their own method as being particularly faithful by optimizing the choice of faithfulness metric (Blücher et al., 2024). It has also been pointed out that XAI methods can be manipulated to yield arbitrary explanations (Dombrowski et al., 2019; Xin et al., 2024). In image prediction tasks, XAI explanations are frequently observed to resemble results of simple edge detection filters (e.g., Adebayo et al., 2018; Kauffmann et al., 2022; Clark et al., 2024). Many XAI methods come in multiple variants, and the criteria for choosing methods and their hyperparameters are often not well justified or documented. Here we argue that a more fundamental limitation of the field – the lack of formal specifications of XAI problems – renders efforts to make explanations more faithful, more robust, more consistent, or more aligned with subjective human judgment, premature. Such efforts will become relevant again once methods that are proven to solve well-posed "explainability problems" in the first place are available. To enable the development of such new methods, the current paradigm of algorithm-driven development needs to be revised. We propose that a new scientific process of XAI development should proceed in six steps: 1. Assessing the use-case-specific information needs of users and stakeholders. 2. Defining the formal requirements and the XAI problems that address these information needs. 3. Designing suitable methods to solve the concrete XAI problems. 4. Performing theoretical analyses w.r.t. the adherence to the formal requirements. 5. Performing empirical validation using appropriate ground-truth benchmarks. 6. Improving the methods with respect to further desiderata.

**Formalizing XAI** It is unreasonable to call a mapping from input features to real numbers an "explanation" without endowing these numbers with a well-defined formal interpretation (e.g. Murdoch et al., 2019). Without a formal problem statement, the ability of an XAI method to answer relevant questions cannot be objectively assessed, and it is not possible to use the method for systematic quality control. Relevant information needs of users may relate to properties of a given ML model, its training data, a given test input, or combinations of these, and may differ between use cases. Additionally, different stakeholders, such as ML developers, users (e.g., physicians or patients), and regulators, have different information needs. Formalizing use-case and stakeholder-specific questions will lead to distinct XAI problems to be addressed by tailored XAI methodologies. Such a formal framework will also provide corresponding provable and testable notions of explanation correctness and quantifiable notions of explanation performance. Previous work (Wilming et al., 2022; Borgonovo et al., 2023) has formulated the presence of a statistical association between feature and target as a minimal requirement for important features. Using this or similar criteria to falsify XAI methods can prevent potentially harmful or costly misinterpretations.

**Theoretical analysis of XAI correctness** Recent insightful theoretical analyses have shown systematic failure modes of XAI methods within respective mathematical frameworks (Sixt et al., 2020; Bilodeau et al., 2024). Once formal notions of explanation correctness become part of such frameworks, XAI methods can be formally analyzed with respect to these. Kindermans et al. (2018) and Wilming et al. (2023) have analyzed popular feature attribution methods and found that many do not fulfill the statistical association property property in the presence of correlated noise. Such research can identify theoretical shortcomings and guide the development of improved methods.

**XAI benchmarking using ground-truth data** Based on formal criteria, it is possible to design ground-truth benchmark datasets that are realistic yet generated from a known and controlled parametric distribution. Various authors have proposed datasets in which the features sharing actual information with the prediction target are known by construction (Ismail et al., 2019; Hooker et al., 2019; Yalcin et al., 2021; Wilming et al., 2022; Arras et al., 2022; Zhou et al., 2022; Clark et al., 2024; Budding et al., 2021; Oliveira et al., 2024; Oramas et al., 2019). This can be used to quantify specific definitions of explanation performance. In Oramas et al. (2019), synthetic image datasets are introduced, where color manipulations of predefined object parts serve as class-related features used as ground-truth explanations. In Wilming et al. (2022) and Clark et al. (2024), a range of popular XAI methods in combination with distinct neural network architectures were benchmarked on linear and non-linear classification problems. In Oliveira et al. (2024), structural magnetic resonance imaging (MRI) data were superimposed with synthetic brain lesions and the effect of pre-training on explanation performance in lesions classification tasks was studied. Wilming et al. (2024) introduced a gender-balanced text dataset and associated gender classification tasks, which allows for quantifying explanation performance and biases in explanations. These datasets are publicly available.

# 6 Conclusions and Outlook

Theoretical and empirical analyses of simple data-generating models have shown that popular feature attribution methods can systematically fail to answer important questions about data and ML models. The main technical limitation of existing methods, causing false interpretations in the considered examples, is the inherent assumption of feature independence. More generally, the field of XAI is impeded by the current paradigm of algorithm- instead of problem-driven development and the lack of formal notions of explanation correctness. These limitations are shared by other XAI paradigms such as concept- or example based explanations. Just as ML in general, though, the field of XAI is fast-developing with novel methodological developments being introduced each year. Recent advancements in algorithmic recourse (Karimi et al., 2020), confounder detection (Janzing and Schölkopf, 2018), and generative modeling (Hvilshøj et al., 2021; Sobieski and Biecek, 2024) promise to address some of the limitations presented here. A systematic formalization and scrutinization of the field of XAI would make it possible to objectively assess the ability of individual approaches to solve specific XAI problems. Researchers should formally define the specific problems that XAI should solve and design methods accordingly. Synthetic data with ground-truth explanations can play an important role in (in)validating XAI methods. This may eventually lead to XAI-based workflows that can indeed be used to systematically provide quality assurance for ML.

**Acknowledgements**

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 758985), by Project 22HLT05 MAIBAI, which has received funding from the European Partnership on Metrology, co-financed from the European Union's Horizon Europe Research and Innovation Programme and by the Participating States, and by the Metrology for Artificial Intelligence in Medicine (M4AIM) programme funded by the German Federal Ministry for Economy and Climate Action (BMWK) in the frame of the QI-Digital Initiative.

## References


Aas, K., Jullum, M., Løland, A., 2021. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. Artificial Intelligence 298, 103502.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Curran Associates Inc.. pp. 9525–9536.

Arras, L., Osman, A., Samek, W., 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion 81, 14–40.

Ates, E., Aksar, B., Leung, V.J., Coskun, A.K., 2021. Counterfactual explanations for multivariate time series, in: 2021 International Conference on Applied Artificial Intelligence (ICAPAI), pp. 1–8.

Babic, B., Gerke, S., Evgeniou, T., Cohen, I.G., 2021. Beware explanations from ai in health care. Science 373, 284–286.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE 10, 1–46.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R., 2010. How to explain individual classification decisions. Journal of Machine Learning Research (JMLR) 11, 1803–1831.

Biessmann, F., Refiano, D., 2021. Quality metrics for transparent machine learning with and without humans in the loop are not correlated, in: ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI. `arXiv:2107.02033`.

Bilodeau, B., Jaques, N., Koh, P.W., Kim, B., 2024. Impossibility theorems for feature attribution. Proceedings of the National Academy of Sciences 121.

Blücher, S., Vielhaben, J., Strodthoff, N., 2022. Preddiff: Explanations and interactions from conditional expectations. Artificial Intelligence 312, 103774.

Blücher, S., Vielhaben, J., Strodthoff, N., 2024. Decoupling pixel flipping and occlusion strategy for consistent XAI benchmarks. Transactions on Machine Learning Research .

Bordt, S., Finck, M., Raidl, E., von Luxburg, U., 2022. Post-hoc explanations fail to achieve their purpose in adversarial contexts, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 891–905.

Borgonovo, E., Ghidini, V., Hahn, R., Plischke, E., 2023. Explaining classifiers with measures of statistical association. Computational Statistics & Data Analysis 182, 107701.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Budding, C., Eitel, F., Ritter, K., Haufe, S., 2021. Evaluating saliency methods on artificial data with different background types, in: Medical Imaging meets NeurIPS. An official NeurIPS Workshop. `arXiv:2112.04882`.

Clark, B., Wilming, R., Haufe, S., 2024. XAI-TRIS: non-linear image benchmarks to quantify false positive post-hoc attribution of feature importance. Machine Learning , 1–40.

Conger, A.J., 1974. A Revised Definition for Suppressor Variables: A Guide To Their Identification and Interpretation. Educational and Psychological Measurement 34, 35–46.

Del Giudice, M., 2024. The prediction-explanation fallacy: A pervasive problem in scientific applications of machine learning. Methodology 20, 22–46.

Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P., 2019. Explanations can be manipulated and geometry is to blame. Advances in neural information processing systems 32.

European Commission, 2021. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts .

Fok, R., Weld, D.S., 2023. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. AI Magazine .

Ghassemi, M., Oakden-Rayner, L., Beam, A.L., 2021. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health 3, e745–e750.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2019. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys 51, 1–42.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110.

Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., 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, 1–11.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9, e1312.

Hooker, S., Erhan, D., Kindermans, P.J., Kim, B., 2019. A benchmark for interpretability methods in deep neural networks, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc.. pp. 9737–9748.

Hvilshøj, F., Iosifidis, A., Assent, I., 2021. Ecinn: efficient counterfactuals from invertible neural networks. arXiv preprint arXiv:2103.13701 .

Ismail, A.A., Gunady, M., Pessoa, L., Corrada Bravo, H., Feizi, S., 2019. Input-Cell Attention Reduces Vanishing Saliency of Recurrent Neural Networks, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.

Jacovi, A., Goldberg, Y., 2020. 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, Association for Computational Linguistics, Online. pp. 4198–4205.

Janzing, D., Schölkopf, B., 2018. Detecting non-causal artifacts in multivariate linear regression models, in: International Conference on Machine Learning, PMLR. pp. 2245–2253.

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., Gama, J., 2021. How can i choose an explainer? an application-grounded evaluation of post-hoc explanations, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 805–815.

Jiménez-Luna, J., Grisoni, F., Schneider, G., 2020. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence 2, 573–584.

Karimi, A.H., Schölkopf, B., Valera, I., 2021. Algorithmic recourse: from counterfactual explanations to interventions, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 353–362.

Karimi, A.H., Von Kügelgen, J., Schölkopf, B., Valera, I., 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. Advances in neural information processing systems 33, 265–277.

Kauffmann, J., Esders, M., Ruff, L., Montavon, G., Samek, W., Müller, K.R., 2022. From clustering to cluster explanations via neural networks. IEEE Transactions on Neural Networks and Learning Systems 35, 1926–1940.

Kindermans, P.J., Schütt, K.T., Alber, M., Müller, K.R., Erhan, D., Kim, B., Dähne, S., 2018. Learning how to explain neural networks: Patternnet and patternattribution, in: International Conference on Learning Representations.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications 10, 1096.

Lundberg, S.M., Lee, S.I., 2017. A Unified Approach to Interpreting Model Predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc.. volume 30, pp. 4765–4774.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. Journal of the Royal Statistical Society Series B: Statistical Methodology 72, 417–473.

Molnar, C., 2020. Interpretable Machine Learning. Independently published.

Montavon, G., Bach, S., Binder, A., Samek, W., Müller, K.R., 2017. Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. Pattern Recognition 65, 211–222.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 116, 22071–22080.

Oliveira, M., Wilming, R., Clark, B., Budding, C., Eitel, F., Ritter, K., Haufe, S., 2024. Benchmarking the influence of pre-training on explanation performance in mr image classification. Frontiers in Artificial Intelligence 7, 1330919.

Oramas, J., Wang, K., Tuytelaars, T., 2019. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks, in: International Conference on Learning Representations.

Rai, A., 2020. Explainable ai: From black box to glass box. Journal of the Academy of Marketing Science 48, 137–141.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " Why should I trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.

Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E., 2022. A consistent and efficient evaluation strategy for attribution methods. `arXiv:2202.00449`.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 206–215.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K., 2017. Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems 28, 2660–2673.

Samek, W., Müller, K.R., 2019. Towards explainable artificial intelligence. Explainable AI: interpreting, explaining and visualizing deep learning , 5–22.

Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q.H., Nguyen, C.D.T., Ngo, V.D., Seekins, J., Blankenberg, F.G., Ng, A.Y., Lungren, M.P., Rajpurkar, P., 2022. Benchmarking saliency methods for chest X-ray interpretation. Nature Machine Intelligence 4, 867–878.

Shapley, L.S., 1953. A value for n-person games. Contributions to the Theory of Games 2, 307–317.

Shmueli, G., 2010. To Explain or to Predict? Statistical Science 25.

Sixt, L., Granz, M., Landgraf, T., 2020. When Explanations Lie: Why Many Modified BP Attributions Fail, in: Proceedings of the 37th International Conference on Machine Learning, PMLR. pp. 9046–9057.

Sobieski, B., Biecek, P., 2024. Global counterfactual directions. arXiv preprint arXiv:2404.12488 .

Sokol, K., Flach, P., 2020. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. KI-Künstliche Intelligenz 34, 235–250.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic Attribution for Deep Networks, in: Precup, D., Teh, Y.W. (Eds.), ICML, PMLR. PMLR. pp. 3319–3328.

Tideman, L.E., Migas, L.G., Djambazova, K.V., Patterson, N.H., Caprioli, R.M., Spraggins, J.M., Van de Plas, R., 2021. Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized shapley additive explanations. Analytica Chimica Acta 1177, 338522.

Ustun, B., Spangher, A., Liu, Y., 2019. Actionable recourse in linear classification, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, Atlanta GA USA. pp. 10–19.

Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech. 31, 841.

Watson, D.S., 2022. Interpretable machine learning for genomics. Human Genetics 141, 1499–1513.

Weber, R.O., Johs, A.J., Goel, P., Silva, J.M., 2024. XAI is in trouble. AI Magazine , aaai.12184.

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., Grosse-Wentrup, M., 2015. Causal interpretation rules for encoding and decoding models in neuroimaging. Neuroimage 110, 48–59.

Wilming, R., Budding, C., Müller, K.R., Haufe, S., 2022. Scrutinizing XAI using linear ground-truth data with suppressor variables. Machine Learning, Special Issue of the ECML PKDD 2022 Journal Track , 1–21.

Wilming, R., Dox, A., Schulz, H., Oliveira, M., Clark, B., Haufe, S., 2024. GECOBench: A gender-controlled text dataset and benchmark for quantifying biases in explanations. arXiv:2406.11547.

Wilming, R., Kieslich, L., Clark, B., Haufe, S., 2023. Theoretical behavior of XAI methods in the presence of suppressor variables, in: Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR, pp. 37091–37107.

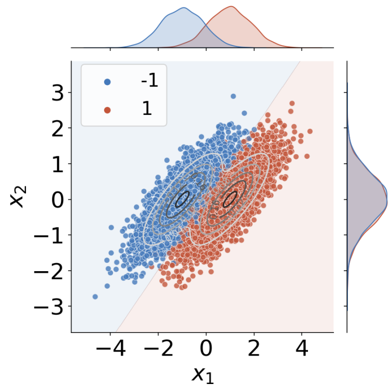
Wong, F., Zheng, E.J., Valeri, J.A., Donghia, N.M., Anahtar, M.N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., Manson, A.L., Friedrichs, J., Helbig, R., Hajian, B., Fiejtek, D.K., Wagner, F.F., Soutter, H.H., Earl, A.M., Stokes, J.M., Renner, L.D., Collins, J.J., 2024. Discovery of a structural class of antibiotics with explainable deep learning. Nature 626, 177–185.

Xin, X., Huang, F., Hooker, G., 2024. Why you should not trust interpretations in machine learning: Adversarial attacks on partial dependence plots. arXiv preprint arXiv:2404.18702 .

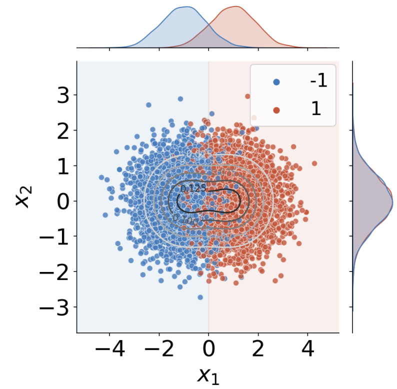
Yalcin, O., Fan, X., Liu, S., 2021. Evaluating the correctness of explainable ai algorithms for classification. arXiv preprint arXiv:2105.09740 .

Zhou, Y., Booth, S., Ribeiro, M.T., Shah, J., 2022. Do feature attribution methods correctly attribute features?, in: Proceedings of the AAAI conference on artificial intelligence. Number: 9.

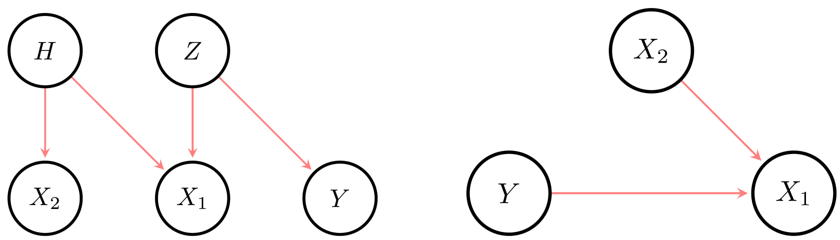
# A Systematic importance attribution to suppressor variables




(a) $c = 0.8$




(b) $c = 0$




(c) Suppressor ($X_2$) and Collider ($X_1$)

Figure 1: a/b) Data sampled from the generative model (Example A) introduced in Section 3.1 (Wilming et al., 2023) for two different correlations $c$ and constant variances $s_1^2 = 0.8$ and $s_2^2 = 0.5$. Boundaries of Bayes-optimal decisions are shown as well. The marginal sample distributions illustrate that feature $X_2$ does not carry any class-related information. c) Causal structure of the data in Examples A (left) and B (right). $X_2$ is a so-called suppressor variable that has no statistical association with the target $Y$, although both influence feature $X_1$, which is called a collider. Figure partially adopted from Wilming et al. (2023).

Table 1: Summary of the results of Kindermans et al. (2018) and Wilming et al. (2023). Various popular feature attribution methods systematically attribute non-zero importance to suppressor variables that have no statistical association to the prediction target. For Shapley values, this property may depend on the chosen value function.

| XAI methods attributing nonzero importance to suppressors |
|---|
| Permutation Feature Importance (Breiman, 2001) |
| Shapley Value (Shapley, 1953) |
| Gradient (Baehrens et al., 2010) |
| LIME (Ribeiro et al., 2016) |
| Faithfulness (Pixel Flipping, Samek et al., 2017) |
| SHAP (Marginal Expectation, Lundberg and Lee, 2017) |
| Counterfactuals (Wachter et al., 2017) |
| Integrated Gradient (Sundararajan et al., 2017) |
| LRP/DTD (Bach et al., 2015; Montavon et al., 2017) |
| Partial Dependency Plot (e.g., Molnar, 2020) |
| SHAP (Conditional Expectation, Aas et al., 2021) |