Correct Explanations and How to Define Them: Properties and Metrics for Measuring Correctness of Three Forms of ML Model Input/Output Behaviour Explanations

VANDITA SINGH*, Ericsson Research, Sweden
KRISTIJONAS ČYRAS, Ericsson Inc., CA, U.S.A & European AI Office, European Commission, Belgium
MUHAMMAD ZAIN AKRAM, Ericsson Research, Sweden
RAFIA INAM, Ericsson Research & Royal Institute of Technology, Sweden

In explainable AI, many explanation methods generate similar yet diverging explanations for machine learning (ML) models. How fair is it then to explain ML model behaviour by such explanations? Arguably, one needs to judge whether those explanations are good at explaining ML model input/output behaviour. We here attempt to formalise ways to judge goodness of such explanations in terms of their *correctness*. For assessing correctness, one needs to have desirable properties of explanation correctness in mind, as well as was to measure satisfaction of those properties. We submit two high-level properties of soundness and completeness for assessing explanation correctness: explaining is sound if the model behaves the way the explanations say; explaining is complete if explanations can be given for model's outputs on any inputs. We formulate soundness and completeness properties for three forms of explanations: feature importance, counterfactuals and rules. We further formalise multiple general metrics, at least one for each property and form of explanation, for quantitatively measuring satisfaction of soundness and completeness. We argue that explanations are correct in as much as various aspects of the different forms of explanations are met as quantified by those metrics. We hope that being able to assess correctness of ML model input/output behaviour explanations against formal properties and metrics is a substantial step towards fairly explaining ML-based inference.

JAIR Track: Fairness and Bias

JAIR Associate Editor: Roberta Calegari

JAIR Reference Format:
Vandita Singh, Kristijonas Čyras, Muhammad Zain Akram, and Rafia Inam. 2025. Correct Explanations and How to Define Them: Properties and Metrics for Measuring Correctness of Three Forms of ML Model Input/Output Behaviour Explanations. *Journal of Artificial Intelligence Research* 84, Article 6 (September 2025), 41 pages. DOI: 10.1613/jair.1.18691

1 Introduction

With the emergence of ever more capable AI-based automated and autonomous systems, concerns regarding safety of and trust in AI systems are growing too. The concerns have escalated to the extent of open letters issued arguing for slowing down if not pausing development of powerful AI systems (Future of Life Institute 2023). Addressing some of those concerns, trustworthy AI (Chatila et al. 2021) concerns creating safe and trustable AI

*Corresponding Author.

Authors' Contact Information: Vandita Singh, ORCID: 0000-0002-0586-8509, vandita.singh@ericsson.com, Ericsson Research, Stockholm, Sweden; Kristijonas Čyras, ORCID: 0000-0002-4353-8121, kcyras@gmail.com, Ericsson Inc., CA, U.S.A & European AI Office, European Commission, Brussels, Belgium; Muhammad Zain Akram, zain-akram@hotmail.com, Ericsson Research, Stockholm, Sweden; Rafia Inam, ORCID: 0000-0001-7448-3381, rafia.inam@ericsson.com, raina@kth.se, Ericsson Research & Royal Institute of Technology, Stockholm, Sweden.

DOI: 10.1613/jair.1.18691

systems. It is a huge umbrella term (see (European Commission and Directorate-General for Communications Networks, Content and Technology 2019)) encompassing many AI research areas, notably AI fairness (Schwartz et al. 2022) and explainable AI (XAI) (Phillips et al. 2021).

AI fairness (also called algorithmic fairness by (Mitchell et al. 2021)) concerns how algorithmic decisions pertain to formalised notions of fair decision making, such as individual fairness of giving similar individuals similar decisions and non-discrimination of giving similar groups on the whole similar decisions (Friedler et al. 2021). Currently, fair AI decision making often pertains to prediction or classification by machine learning (ML) models. At the same time, XAI concerns making AI systems intelligible (Adadi and Berrada 2018). Currently, explainability research often focuses on explaining internal workings and input/output behaviour of ML models. XAI and AI fairness are not only two complementary areas of trustworthy AI, but could sometimes even be seen as two sides of the same coin. For instance, Ignatiev, Cooper, et al. (2020) consider how explanations of ML model input/output behaviour can serve to assess fairness thereof. We take a complimentary viewpoint and try to address the question, loosely stated, as to whether it is fair to explain the behaviour of an AI system/ML model using the (by now) common kinds of explanations.

To exemplify, there often are various explanation methods, henceforth called *explainers*, that produce explanations of the same kind for a given ML model and its input instances. For instance, an attributive explanation method such as SHAP (SHapley Additive exPlanations) (Lundberg and Lee 2017) assigns *feature importance* scores to the input features for a particular prediction made by a given model. Another explanation method, such as Integrated Gradients (IG) (Sundararajan et al. 2017), also attributes feature importance scores to features for the same prediction by the same model. If the feature importance attributions differ, as they often do, how is it fair to use one explainer or another to explain the model's behaviour? This challenge to decide which of the diverging explanations are better or appropriate is generally known as the disagreement problem among explanations (Agarwal et al. 2022; Krishna et al. 2022).

In some cases, such as with attributive explainers, they may work in fundamentally distinct ways – combinatorial attribution in SHAP and integration in IG – even if they generate explanations of the same *form*. In other cases, such as with *counterfactual* explanations, the explainers generate explanations of the same form but typically by optimising different targets or constraints. With explanations in the form of *rules*, explainers may work heuristically or provide guarantees of some aspects of explanations, such as their fidelity to the underlying model. In any case, explanations of the same kind and form are typically intended to carry the same meaning, whether feature importance, what-if scenarios (counterfactuals) or if-then implications (rules). Whether or not the disagreement problem arises, the explanations should be judged according to how good or appropriate they are, e.g. how correct, robust, efficient, intelligible, or whatever the intended measure(s) (Kaur et al. 2022). In the absence of such measures of goodness, the intended receiver of explanations might face challenges to perform a fair selection among explanations or explainers, and may thus partake in an unfair explaining process.

We here seek to address the challenge of assessing the goodness of explanations for the sake of explaining fairly. Fundamentally, we posit that it is imperative for explanations to be *correct* in order for the explaining to be fair. We do not claim that it is sufficient for explanations to be correct, but rather argue that it is necessary for explanations to meet some measurable standards of correctness if they are to explain ML models. To define what explanation correctness means, we get inspired (like (Kulesza et al. 2013) do) by the notions of sound and complete reasoning in formal logics. Colloquially, soundness demands that if one can prove something, then that thing is true. *Vice versa*, completeness demands that if something is true, then one can prove it. We attempt an intuitive analogy for sound and complete explaining (in the spirit of (Čyras, Letsios, et al. 2019)). *Soundness* would require that if there is an explanation, then the model actually behaves in accordance with it. *Completeness* would require that if the model behaves in some way, then there is an explanation for that. As much as explanations of ML models are sound and complete, we deem them correct. And we believe it is only fair to give correct explanations so as the explaining instills trust in the AI system.

In this paper we consider the problem of evaluating the *correctness* of explanations generated by different XAI methods. Specifically, we posit that explanation correctness amounts to satisfaction of two desirable *properties of explanations*, namely soundness and completeness. We further submit that satisfaction of (such or similar) properties of explanations should be quantified using *metrics*. We attempt to formally specify properties and metrics for assessing correctness of explanations in certain settings on AI decision making. Concretely, we focus on ML model input/output behaviour explanations and consider three forms of such explanations with respect to ML models trained for classification tasks on tabular data. The three forms of explanations are feature importance-based, rule-based and counterfactual explanations. We formally define explanations as mathematical artefacts that are meant to capture a range of such explanations generated by different explainers. We further postulate desirable correctness properties and a multitude of metrics thereof for each of the three forms of explanations. The forms of explanations and their soundness and completeness can be intuitively stated as follows.

A rule-based explanation is an if-then rule with premises consisting of feature-value pairs and conclusion consisting of the predicted class. We will deem a rule sound in as much as its conclusion agrees with the model output for inputs covered by the premises of the rule; and complete in as much as it covers the model's inputs/outputs and is representative of rules that so explain the model. A counterfactual explanation is an input instance consisting of feature-value pairs. We will deem counterfactual explanations to a given instance sound in as much as the counterfactuals are possible and have the model's outputs differ from the one for the explained input; and complete in as much as they are diverse and cover model inputs. A feature importance-based explanation is a set of feature importance scores. We will deem feature importance scores sound in as much as they adequately reflect the impact that features have on the model's outputs; and complete in as much as they are representative of such impact scores. We will see how different metrics can be used to quantify the conceptual aspects of these formulations.

With this paper, we thus aim to address the following **research question**.

> Can we formulate desirable properties of, and formalise metrics for, assessing explanation correctness of different forms of explanations of the input/output behaviour of ML models trained for classification tasks on tabular data?

We answer in the affirmative by formulating the properties of soundness and completeness of explanations and formalising metrics for measuring those properties independently of the explanation method used for generating the explanations. We do this for three forms of explanations: rule-based, feature importance-based and counterfactuals.

In brief, our **main contributions** are the following. We state six formal properties of explanation correctness, three of soundness and and three of completeness, one for each form of explanations. Colloquially, soundness mandates that explanations should be truthful with respect to the ML model explained, and completeness mandates that explanations should generalise to cover the model behaviour. We follow up with 12 metrics to quantify satisfaction of the properties (discussed in detail in Section 4, with a schematic in Figure 1 introduced therein), with the following origins:

- five metrics are *adoptions* from existing literature with minor modifications – namely, 4.1.1. Fidelity (rule soundness), 4.1.2. Coverage (rule completeness), 4.2.1. Validity and 4.2.1. Plausibility (counterfactual soundness), and 4.2.2. Diversity (counterfactual completeness);
- four are *generalisations* – namely, 4.2.1. Feasibility (counterfactual soundness), 4.3.1. Fidelity and 4.3.1. Validity (feature importance soundness), and 4.3.1. Agreement (feature importance soundness);
- three are *new* completeness metrics – namely, 4.1.2. Representativeness (rules), 4.2.2. Coverage (counterfactuals) and 4.3.2. Representativeness (feature importance).

We exemplify our framework for assessing explanation correctness with hand-crafted examples as well as with illustrative experimental results using six off-the-shelf ML model explainers with boosted tree ensemble and neural network models trained on three datasets.

In this work we take the *functionally-grounded assessment* view of explainers and explanations (Doshi-Velez and Kim 2017; Nauta et al. 2023) in that *we focus on functional (but potentially user-informed) metrics rather than user evaluation studies*. We further concur with (Nauta et al. 2023) to address only those aspects of explanations that "*directly* influence explanation quality" (Nauta et al. 2023, p. 10, emphasis original), particularly ignoring computation resources required to generate explanations or presentation of explanations. In particular, we will focus on correctness of explanations and how it can be characterised via the proposed properties of soundness and completeness, to be assessed using functional metrics that directly pertain to quality of explanations as computational artefacts.

In what follows we first discuss related work, mostly on assessment of (goodness of) explanations. We then give preliminaries for our work: we define the forms of explanations, discuss the notions of metrics and properties, and give the setup for basic empirical illustrations that will accompany the formal exposition of explanation correctness. We then detail our take on explanation correctness via properties of soundness and completeness as measured by the various metrics proposed for assessing explanations. We finally discuss in brief the main theoretical contribution of this paper, namely the formulation of correctness properties (that we believe are necessary for the explaining process to be fair) and the formalisation of the metrics for quantitatively assessing those properties with respect to three different forms of common explanations of ML model input/output behaviour.

2 Related Work

XAI has grown enormously over the past few decades and there are now a plethora of explainability techniques, explainers, and kinds of explanations concerning numerous kinds of AI systems, including symbolic, neuro-symbolic, and statistical models – see (A. and R. 2023; Anjomshoae et al. 2019; Arrieta et al. 2020; Burkart and Huber 2021; Carvalho et al. 2019; Chakraborti et al. 2020; Ciatto, Sabbatini, et al. 2024; Čyras, Badrinath, et al. 2021; Čyras, Rago, et al. 2021; Ding et al. 2022; Fandinno and Schulz 2019; Graziani et al. 2023; Ibrahim and Shafiq 2023; Linardatos et al. 2021; Nunes and Jannach 2017; Ras et al. 2022; Schwalbe and Finzel 2023; Swartout et al. 1991) for overviews. There are numerous works concerning qualitative and quantitative assessment of explainers and explanations using properties and metrics, typically with respect to ML models – see (Agarwal et al. 2022; Arya, Bellamy, P. Chen, et al. 2019; Bhatt et al. 2021; Carvalho et al. 2019; Doshi-Velez and Kim 2017; Guidotti 2022; A.-H. Karimi et al. 2020; Kulesza et al. 2013; Laugel et al. 2019; Nauta et al. 2023; Robnik-Šikonja and Bohanec 2018; Rosenfeld 2021; Samek et al. 2016; Sokol and Flach 2020; Zhou et al. 2021). We will discuss the most specific ones, particularly (Guidotti 2022; Kulesza et al. 2013; Nauta et al. 2023), that we borrow from or build upon in Section 4. Here, we briefly discuss those works as well as more generally the works that collect lists of properties and metrics for assessing quality of explanations, namely (Carvalho et al. 2019; Hedström et al. 2023; Robnik-Šikonja and Bohanec 2018; Sokol and Flach 2020; Zhou et al. 2021).

The early work of Kulesza et al. (2013) on assessing goodness of ML model explanations introduced the viewpoint of explanation soundness and completeness, borrowing the notions from formal logics. In a nutshell, the idea behind these two properties is that an explanation is sound in as much as the model behaves the way the explanation says, and complete in as much as it explains all of the model. That work used user studies to empirically evaluate how well explanations of different degrees of objectively measured soundness and completeness correspond to user perceived reasons behind algorithmic recommendations of music songs. They concluded that while aiming for sound and complete explanations is a good design choice when developing explainers, it may bring unwanted costs in cases where the audience is unwilling to attend to the explanations.

We instead focus on formalising the two properties of explanation soundness and completeness for three forms of explanations, arguing for desirability of their satisfaction and assigning computational metrics to measure it. Nonetheless, we are greatly inspired by Kulesza et al. (2013) in our conceptual understanding of explanation correctness in terms of soundness and completeness.

Robnik-Šikonja and Bohanec (2018) is an early work to state various properties of ML model explainers and explanations thereof, though not entirely formally or unambiguously (as noted in e.g. (Carvalho et al. 2019)). The several properties of 'explanation quality' summed in (Robnik-Šikonja and Bohanec 2018) are fidelity, accuracy, representativeness, certainty, novelty, stability, consistency, comprehensibility, importance. Fidelity, with both local and global versions, pertains to how well an explanation reflects the behaviour of the underlying ML model, locally or globally. Relatedly, accuracy pertains to how well an explanation generalises to multiple data instances and their predictions and representativeness to how well explanations cover the model's behaviour. We will be inspired by all these three in our formulations of explanation correctness. Linked to the three above are certainty and novelty, which respectively pertain to whether an explanation reports the model's confidence and whether the data instance is from a well-represented data region. These we deem irrelevant to correctness of explanations, because they concern the model and data characteristics, respectively, rather than the explanation directly.

In terms of comparing explanations themselves, stability pertains to the similarity of explanations generated for similar instances while consistency pertains to the similarity of explanations generated for different models (for the same task). We are not interested in similarity of explanations given different ML models, as we deem it to be very use case-specific to decide whether explanations should be similar across (dis)similar models. We submit these aspects reflect the more general property of *robustness* of explanations. It is widely considered and measured via different metrics (and properties) in many works. For instance, sensitivity (Bhatt et al. 2021; Sundararajan et al. 2017) measures an explanation's ability to discover sensitive aspects of the model when subjected to perturbations. Similarly, consistency measures if explanations correspond to the same inputs and outputs. Relatedly, (in)stability measures relative changes to an explanation with respect to changes in model inputs or performance (Agarwal et al. 2022). Particularly in case of counterfactual explanations, stability could be measured by estimating whether the counterfactuals are close enough in case of slight perturbations to the input instances (Albini et al. 2022; Guidotti 2022; Laugel et al. 2019; Upadhyay et al. 2021). These and similar notions are grouped under the 'continuity' property in (Nauta et al. 2023). We do not engage with any of these aspects of explanations, since we generally consider robustness to be tangential to correctness.

Regarding comprehension aspects of explanations as in (Carvalho et al. 2019; Robnik-Šikonja and Bohanec 2018), comprehensibility pertains to understandibility of explanations, often judged in terms of explanation size and artefacts therein (such as features or rules). Importance further details if an explanation reports any degrees of importance for its artefacts, such as feature scores or rule weights. Comprehension is also not part of correctness in our opinion, and is also clearly audience-dependent.

We depart from both Robnik-Šikonja and Bohanec (2018) and Carvalho et al. (2019) in that we consider some of the above aspects as metrics to measure explanations, but think of properties at a higher level. For instance, fidelity (which is essentially 'correctness' in (Carvalho et al. 2019)) measures soundness of feature importance-based explanations, whereas accuracy and representativeness measure completeness of the same. What Carvalho et al. (2019) call 'completeness' will for us essentially amount to coverage of explanations. We will consider soundness and completeness as two higher level properties, with different formulations and metrics for different forms of explanations. As noted in (Carvalho et al. 2019), it is not clear how to formalise or measure the above discussed aspects of explanations, which is something that we tackle in this paper for the aspects that we deem relevant to explanation correctness.

The authors of (Robnik-Šikonja and Bohanec 2018) further state four properties of explainers – namely, expressive power, translucency, portability and algorithmic complexity – which can be summarised thus. Expressive power refers to the language or structure of the explanations, e.g. feature importance scores or if-then form-of

rules. Translucency and portability refer to, respectively, whether the explainer needs access to the ML model internals and with what kinds of ML models it can work. We are agnostic with respect to both; we will only care about how to evaluate explainers as long as explanations are generated, given some underlying ML model. The same indifference applies to algorithmic complexity in our case. So none of the above we consider in this work as properties against which to evaluate explainers but rather as characteristics of explainers or their applicability.

The authors of (Carvalho et al. 2019), building on (Miller et al. 2017), also list several human-centric characteristics of explanations, namely that they should be contrastive, selective, social, focusing on the abnormal, truthful, general and probable as well as consistent with beliefs. The first four seem for us to be characteristics of explainers or their applicability, not about correctness of explanations. Being truthful, general and probable, in the sense of applying to various ML model predictions and speaking truth about them, is something that we will consider as parts of soundness and completeness. They are in any case not formalised in any way in the above works, whereas we will attempt that. Lastly, consistence with prior beliefs is very much audience/user-centric, a type of aspect outside the scope of this paper.

In the "explainability fact sheets" paper (Sokol and Flach 2020), the authors gathered a collection of requirements for assessing explainers and explanations. They presented five families of requirements, namely functional, operational, usability, safety and validation. Functional ones consider algorithmic requirements on the explainer, such as type of ML model or task. These do not pertain to directly assessing the quality of explanations and for us amount to explainer applicability aspects. The operational requirements mostly overlap with our forms of explanations and the model setup. Safety requirements pertain to robustness of explanations and explainers – again, out of our scope here. Validation pertains to empirical evaluation of effectiveness of the explanation process using either user studies or synthetic scenarios, likewise out of scope of this work. Finally, even though usability is mostly user-centric according to the authors, it includes aspects that we deem relevant to evaluating explanation correctness. In particular, the authors consider soundness and completeness as two requirements that would measure truthfulness and generalisation, respectively, of explanations. We do get inspired by their conceptual use of the notions of completeness and soundness, but depart from them by focusing deeply on the two properties, formulating them properly for different forms of explanations and formalising metrics to measure their satisfaction. We however ignore the other usability requirements, namely contextfullness, interactiveness, actionability, chronology, coherence, novelty, complexity, personalisation and parsimony, as those do not pertain to correctness the way we see it.

Zhou et al. (2021) presents an overview of axiomatic properties and quantitative metrics for assessing ML model explanations. They broadly consider two aspects of explainability, namely 'interpretability' and 'fidelity'. The first one pertains to explanations being understandable to a human user and is characterised using three properties: clarity, broadness and parsimony. Of these, broadness pertains to how general an explanation is, and perhaps conceptually falls under what we consider completeness of explanations. Their fidelity, on the other hand, essentially amounts to our correctness, and is in fact characterised using two properties, namely soundness and completeness. They are conceptually the same as those of (Kulesza et al. 2013), and hence similar to what will be our properties of soundness and completeness. The authors of (Zhou et al. 2021) also consider what they call three types of explanation methods, namely model-based explanations, attribution-based explanations and example-based explanations. These are similar to the three forms of explanations – rules, feature importance, counterfactuals – that we will consider, but could be seen to be more general: model-based explanations aim to explain a given ML model using a simpler, interpretable model, so a rule-based model could potentially count as one; counterfactuals are a form of example-based explanations; and attribution-based explanations essentially amount to feature importance. The authors present a taxonomy of quantitative metrics for measuring their properties of explanations for each type of explanation.

We note that the above work, similarly to most others, is largely a classification and collection of prior works, with vague and informal formulations of properties and references to instances of metrics. We instead offer a

more critical reflection of the properties of interest, namely soundness and completeness for assessing explanation correctness, formulate them for three well-defined forms of explanations, as well as argue why they are desirable and how to measure their satisfaction using formalised general metrics that include as instances some of those found in prior art. Nonetheless, Zhou et al. (2021) includes several metrics, notably for measuring soundness of feature importance-based explanations, that are covered in our work. We also note that their classification suggests only a single instance of a completeness metric, specifically for counterfactuals, and barely one metric for soundness of model/rule-based explanations (namely '(dis)agreement' (Lakkaraju et al. 2019), which we will call fidelity in Section 4.1.1). In connection to this, we note that the recent work (Agarwal et al. 2022) to consider the disagreement problem (Krishna et al. 2022) among explanations – diverging explanations generated by different explainers for the same thing – includes several metrics for evaluating faithfulness of attribution-based explanations, notably those for counting (dis)agreement between explanations. Relatedly, Hedström et al. (2023) presents a taxonomy (and an implementation toolkit) of metrics applicable to attribution-based explanations, notably with a collection of metrics in prior works for measuring faithfulness. We will use some of the ideas from the above works for measuring soundness of feature importance-based explanations in this work. However, we will offer more; importantly, a different, formalisation-driven perspective on properties and metrics for the three forms of explanations.

Yet another broad survey of XAI evaluation techniques covering quantitative and qualitative measures is presented in (Nauta et al. 2023). It is a comprehensive classification of 13 properties and various metrics for functionally measuring those properties of ML model explanations, 7 of which concern the content of explanations while others concern either presentation or user-related aspects. They authors there provide a valuable classification of explanation metrics, but do not make an attempt to describe in detail or formalise them, instead giving references to examples of such measures in the literature. We build upon and discuss this work in more depth throughout Section 4.

3 Preliminaries: Notions of Explanation, Metrics and Properties

To assess explainability of an AI system, one analyses explanations obtained from any one explanation method (henceforth, *explainer*) and potentially compares explanations obtained from different explainers (Agarwal et al. 2022; Ciatto, Schumacher, et al. 2020; Krishna et al. 2022; Nauta et al. 2023; Singh 2021). To this end, it is instructive to quantify various aspects of an explanation and to infer characteristics of explanations or explainers, similarly to how one measures the quality of the outputs of an ML model (or any automated system, for that matter). To quantitatively evaluate explanations produced by various ML model explainers, one typically has in mind some desirable *properties* or characteristics of explanations. For instance, it is often desirable that an explainer generates 'correct' explanations in terms of how faithfully they track model's inputs/outputs. In this section, we introduce the notion of *metrics* for assessing explanations, as well as the notion of (more or less desirable) properties of explanations that metrics allow to asses. Before that though, we give the setting of explanations.

3.1 Explanations

In this work we study **explanations of input/output behaviour**, typically of differentiable or tree-based ML models **trained for classification tasks on tabular data**. We focus on ways to assess *post hoc* explanations generated by an explainer after the trained ML model has provided an output (i.e. classification or prediction – we use these terms interchangeably). We consider explanations of both **local and global scope** – the former explain the ML model output for a particular input, while the latter explain the model's overall behaviour, for any input. For example, a global attributive explanation may indicate how much impact each input feature has on model predictions, in aggregate; a local counterfactual explanation, on the other hand, may pinpoint the closest instance that is similar to the input but results in a different prediction.

To set some notation, we consider a fixed but otherwise arbitrary ML model $\mathcal{M}$ trained on a dataset X with feature vector $\mathbf{X}$. The model takes as input an instance $\mathbf{x} \in X$ and outputs $y \in Y$ from data class labels Y. We then consider a likewise fixed but otherwise arbitrary explainer EX that explains the input/output behaviour of $\mathcal{M}$. At a high level, we say that a local-scope explanation $E_{\mathbf{x}}$ is generated by an explainer EX for the model $\mathcal{M}$ with input $\mathbf{x} \in X$ from (some part of) the dataset X and output $\mathcal{M}(\mathbf{x}) = y \in Y$, namely

$$\mathsf{EX}\,(\mathcal{M}(\mathbf{x}) = y) = E_{\mathbf{x}}. \qquad \text{(Local explanation)}$$

Similarly, a global-scope explanation $E_{\mathbf{X}}$ is generated by an explainer EX for the model $\mathcal{M}$ and input feature vector $\mathbf{X}$ (possibly with some concrete feature values), namely

$$\mathsf{EX}\,(\mathcal{M}, \mathbf{X}) = E_{\mathbf{X}}. \qquad \text{(Global explanation)}$$

In general, we refer to either a local- or global-scope explanation as simply E.

In this work we consider **three forms of explanations**. *Intuitively*, those are as follows:

(1) **feature importance** – e.g. a feature importance vector indicating how each feature $f \in \mathbf{X}$ from the input feature vector $\mathbf{X}$ impacts $\mathcal{M}(\mathbf{x}) = y$ (local scope), or how each feature $f \in \mathbf{X}$ impacts $\mathcal{M}$ in aggregate (global scope);
(2) **rules** – e.g. if-then rule comprising of feature-values of $\mathbf{x}$ that are minimally sufficient for $\mathcal{M}$ to output y (local scope), or of feature-value sets that are collectively sufficient for $\mathcal{M}$ to yield output y on input $\mathbf{x}$ whenever $\mathbf{x}$ is covered by the rule (global scope);
(3) **counterfactuals** – e.g. input instance $\mathbf{x}'$ most similar to $\mathbf{x}$ but with $\mathcal{M}(\mathbf{x}') \neq y$ (local scope only, since counterfactuals do not have the aspect of globality).

We *formally* assert the three forms of explanations as follows.

Definition 3.1 (Forms of explanations). Let the labeless dataset $X \subseteq \prod_{i=1}^{|\mathbf{X}|} D_i$ have domain the Cartesian product of feature domains D_i and let Y be the set of class labels of X. We say a local- or global-scope **explanation** E generated by an explainer EX **is**

- **feature importance-based** if it is expressed as a set $\{f_1 : s_1, \ldots, f_n : s_n\}$ of feature-importance score pairs, where $f_j \in \mathbf{X}$ is a feature and $s_j \in \mathbb{R}$ is its importance score, for $j \in \{1, \ldots, n\}$ and $n \leqslant |\mathbf{X}|$ (not all features need be included);
- **rule-based** if it is expressed as a pair[1] $(\{f_1 : s_1, \ldots, f_n : s_n\}, c)$ with the antecedent (also called premises) $\{f_1 : s_1, \ldots, f_n : s_n\}$ comprising feature-value pairs and consequent (also called conclusion) being class label $c \in Y$, where $f_j \in \mathbf{X}$ is a feature and $s_j \subseteq D_j$ is a set of values[2] from the domain D_j, for $j \in \{1, \ldots, n\}$ and $n \leqslant |\mathbf{X}|$;
- **a counterfactual** if it is expressed as an instance[3] $\mathbf{x}' = (f_1 : v_1, \ldots, f_{|\mathbf{X}|} : v_{|\mathbf{X}|})$, where $f_j \in \mathbf{X}$ is a feature and $v_j \in D_j$ is its value from the feature domain D_j, for $j \in \{1, \ldots, |\mathbf{X}|\}$.

We note that a similar yet informal definition of feature importance-based explanations is given in (Yeh 2019, p. 2) while similar definitions of counterfactual and rule-based explanations can be found in, respectively, (Guidotti 2022) and (Lakkaraju et al. 2019). We do not present this definition as a contribution, but rather as formalisation of preliminaries.

With Definition 3.1 we are not aiming to be all-encompassing in the sense of capturing all existing variations of each of the forms of explanations above. For example, feature importance-based explanations may also carry uncertainty estimates for each feature-value pair (see e.g. (Shaikhina et al. 2021)) and rule-based explanations

[1] Possibly a non-empty set $\{R_1, \ldots, R_m\}$ of such pairs, each named R_k.
[2] If a single value v_j is of interest, we can without the loss of generality consider the singleton $s_j = \{v_j\}$.
[3] Possibly a non-empty set $\{\mathbf{x}'_1, \ldots, \mathbf{x}'_m\}$ of such instances.

may carry precision or coverage values (e.g. (Ribeiro et al. 2018)), neither of which are part of our definition. Counterfactuals are also often expressed via only feature-value pairs that differ from the original instance, leaving the unchanged ones implicit. However, it should be straightforward to extend the definition if needed; we simply aim at capturing the representative forms of explanations.

We are not exclusive with the definition either, in not including other forms of explanations, such as graph- or model-based (see e.g. (Čyras, Rago, et al. 2021; Nauta et al. 2023)). We are purposefully limiting the scope of the paper to the three forms of explanations.

We invite the reader to consult the detailed survey of explainers and forms of explanations in (Bodria et al. 2023), especially Fig. 1 on p. 1724 for a visual illustration of their taxonomy of explanations with examples. Henceforth, we will typically refrain from referring to any particular explainers that could generate explanations of a given form, unless they are not mentioned among explainers for tabular data in (Bodria et al. 2023, Table 1, p. 1728). We give our illustration of how such explanations could look like for a toy model next.

Example 3.2. Consider model $\mathcal{M}$, for example a decision tree, that learnt the logical OR function of two binary variables, i.e. $\mathcal{M}(x_1, x_2) = x_1 \vee x_2$ where $(x_1, x_2) \in X = \{0, 1\}^2$ range over values of the feature vector $\mathbf{X} = (f_1, f_2)$.

Rule-based explanations could be the following:[4] $\mathcal{R} = \{R_1 = (\{f_1 : 1\}, 1), R_2 = (\{f_2 : 1\}, 1), R_0 = (\{f_1 : 0, f_2 : 0\}, 0)\}$. Here, R_1 says it is sufficient to set the first variable (that assigns value to feature f_1) to 1 in order to get output class 1; likewise R_2 for the second variable. Meanwhile, R_0 says both variables have to be 0 to yield 0. Without a fixed instance, these are global-scope explanations. But they could well be of local scope too, for example with the instance $\mathbf{x} = (1, 0)$ the rule R_1 says that it suffices to have f_1 value 1 for $\mathcal{M}$ to classify $\mathbf{x}$ as 1.

Counterfactual explanations for the input instance $\mathbf{x} = (0, 0)$ could be the following: $C = \{\mathbf{x}'_1 = (f_1 : 1, f_2 : 0), \mathbf{x}'_2 = (f_1 : 0, f_2 : 1), \mathbf{x}'_3 = (f_1 : 1, f_2 : 1)\}$. Here, $\mathbf{x}'_1$ is to be interpreted as a counterfactual to $\mathbf{x}$ where if only f_1 value were 1, then the model's output would be different: $\mathcal{M}(\mathbf{x}'_1) \neq 0 = \mathcal{M}(\mathbf{x})$. Similarly for $\mathbf{x}'_2$. Whereas $\mathbf{x}'_3$ indicates that if only both f_1 and f_2 values were 1, then the output would be different.

Finally, a (local or global) feature importance-based explanation could be $E = \{f_1 : 1/2, f_2 : 1/2\}$. It expresses that both features are equally important for $\mathcal{M}$ to yield its output (for either any specific instance, or in aggregate).

We finally note that we restrict our attention to classification tasks with tabular data mainly for ease of exposition and intuitive appreciation of the framework to be presented. We immediately observe that extension to regression tasks seems achievable, with minimal changes in formal details, such as letting Y to be class label probabilities, and modifying the metrics accordingly. (Since, as we will see, our metrics typically work with class labels, with regression (as opposed to classification) the comparison to model outcomes could be taken as the class probability difference or (dis)agreement on the class with highest probability score.) It also seems that the setup could be extended to Y being some other co-domain such as the space of data clusters, vocabulary of tokens or a set of actions, thus with a possibility to adapt to self-supervised generation, unsupervised clustering or even (forms of) reinforcement learning, as long as there are explainers supporting such models. Other data modalities, such as images or language, also seem within grasp. However, we refrain from claiming any such generality of our framework, focusing on the generally familiar and approachable setting of classification with tabular data. We believe this setting to be well-suited for understandable formalisations of properties and metrics, which is our main goal, even if at some expense of generality and applicability. We hope that the readers will be able to extend or adapt our framework to other settings.

Having in mind the forms of interest that explanations of ML models trained on tabular data for classification tasks take, we next introduce at a high level the notion of metrics for assessing explanations.

[4]These could be generated by an explainer that returns for $\mathcal{M}$ the $\subseteq$-minimal rules that are sufficient to classify all four inputs. An examplary such explainer is detailed in (Ignatiev, Izza, et al. 2022), which maps a decision tree into propositional logic and returns rule-based explanations in the form of logical implications.

3.2 Metrics

A metric is typically a function that takes parameters of an AI system and yields quantitative values that measure the system's performance in terms of a characteristic of interest. When explanations are generated by an ML model explainer, a metric's value is computed by applying some function to the explanations represented by e.g. input feature importance scores, rules or data instances. For example, fidelity of either a local- or global-scope feature importance-based explanation may be measured by the effect on correlation between input instance perturbations weighted by the feature importance scores and the shift in model outputs for those perturbed inputs; diversity of counterfactual explanations may be measured by aggregating distances among counterfactuals, with, say, the Hamming distance giving the number of features with diverging values between two data instances; fidelity of a local/global rule-based explanation may be measured as the ratio of instances covered by the explanation with the model output matching the rule conclusion against all the instances covered by the explanation.

For our purposes, at a high level, given a global-scope explanation $E_{\mathbf{X}}$ generated by an explainer EX, for a model $\mathcal{M}$ trained on (some part of the) dataset X with input feature vector $\mathbf{X}$, a metric Q is measured by assigning to $E_{\mathbf{X}}$ a numerical value

$$Q_{\mathsf{EX}}(\mathcal{M}, \mathbf{X}, E_{\mathbf{X}}) \in \mathbb{R} \tag{1}$$

Given a local-scope explanation $E_{\mathbf{x}}$ generated by EX, with the model input $\mathbf{x} \in X$ and output $\mathcal{M}(\mathbf{x}) = y \in Y$, a metric Q is measured by assigning to $E_{\mathbf{x}}$ a numerical value

$$Q_{\mathsf{EX}}\left(\mathcal{M}, X, \mathbf{x}, y, E_{\mathbf{x}}\right) \in \mathbb{R} \tag{2}$$

Henceforth, when no ambiguity arises, we may write simply $Q(E)$, for E being either $E_{\mathbf{x}}$ or $E_{\mathbf{X}}$, to denote the value assigned to either local- or global-scope explanation E by metric Q.

With concrete metrics the value of Q may depend on a number of parameters, such as selection of explanations, constraints on feature values, choice of distance functions or reference sets. We will see concrete examples of explanation metrics and ways to measure them in Section 4. Before that, let us talk about how explanation metrics relate to what we call properties of explanations.

3.3 Properties

Properties (Orilia and Paolini Paoletti 2022) refer to characteristics of an object or a system that are typically conceptual to begin with, and are progressively derived from qualitative analysis or quantitative measurements of the object/system. We consider properties of explanations to be a higher level concept than metrics, often encompassing multiple metrics to measure explanations from either the same or different explainers. In this, we concur with (Nauta et al. 2023) that evaluating explanations and explainers necessitates a multi-dimensional view of assessing to what extent various properties are satisfied as measured by different metrics.

We are interested in evaluating the aspect of *explanation correctness*. Intuitively, correct explanations truthfully explain model's behaviour. We will submit two concrete properties of **soundness** and **completeness** for characterising this intuitive notion of correctness with respect to the three forms of explanations. We will formulate the two properties of soundness and completeness as conceptually desirable characterisations of explanations. We will formulate these properties in natural language, to be more or less prescriptive about what each of the form of explanations should satisfy to be correct. We will consider various metrics for each of the three forms of explanations and will use those metrics to assess soundness and completeness of explanations. Essentially, we will deem an explanation either sound or correct in as much as it satisfies some desiderata (e.g. being faithful) as measured by the relevant metrics (e.g. fidelity).

3.4 Empirical Illustration Setup

To illustrate the applicability of our formulations and formalisations to explanations produced by existing explainers, we in some cases use empirical examples. In particular, we will generate some feature importance-based, counterfactual and rule-based explanations to illustrate several instantiations of metrics, where an instantiation amounts to a choice of specific mathematical functions, such as distance or similarity measures. We will also give some examples by-hand, where it is easy to appreciate the application of a metric to an explanation without aggregating over a set of explanations. We here state our empirical setup (as a considerable variation and extension of experiments by (Singh et al. 2022)) for generating examples of explanations. In short, we experiment (in Python 3.11.8) with three tabular datasets (two common ones and one proprietary), six ML classification models (tree-based and neural networks), and six explainers – three for feature importance, two for counterfactuals, one for rules.

We use the well-known and publicly available dataset *Iris* (R. A. Fisher 1936), a locally generated synthetic dataset of the well-known *Spiral* type and a proprietary but publicly discussed dataset called *Telecom* (Terra et al. 2020). Both Iris and Spiral are loaded via scikit-learn 1.6.1 (Pedregosa et al. 2011), the latter using the make_moons method for 100 samples with noise=0.2 (here and henceforth, we specify only the parameters relevant for reproduction rather than replication, thus omitting ones such as random_state=42). Telecom dataset consists of time-series data generated using a telecommunications network simulator: it contains network measurement samples characterised by 12 real-valued features and a target real-valued feature representing network latency. The series is turned into a tabular dataset with a binary classification as to whether three consecutive samples exhibit network latency over a specified threshold. Due to high computation costs of generating explanations using some explainers, we restrict the dataset to 1000 samples. All three datasets are scaled using sklearn.preprocessing.StandardScaler, and 80%/20% train/test split, resulting into X_{train}/X_{test} set sizes of 90/30, 75/25 and 800/200 for Iris, Spiral and Telecom, respectively. We refer the reader to the cited sources for dataset details, but note that they do not matter for our purposes, as long as we can train ML models on the datasets and apply explainers thereafter.

We opt for two well-known kinds of ML models. We train boosted trees via XGBoost (T. Chen and Guestrin 2016) xgboost 2.1.3 (XGB in short) for each dataset, with 100 trees (n_estimators) for Iris and 1000 for Spiral and Telecom, and maximum tree depth max_depth=10, learning_rate=0.1 and eval_metric=logloss. We also train neural networks (NNs) of fully-connected Multi-Layer Perceptron (MLP) type, with two hidden layers of size 16, Rectified Linear Unit (ReLU) activation functions in between and softmax on the output, via PyTorch (Paszke et al. 2019) torch 2.5.1. We train for epochs=10 with batch_size=32 and learning_rate=0.01, optimising using Adam (Kingma and Ba 2015). All models are built with otherwise default parameters, not thoroughly optimised for predictive performance (with XGB and NN accuracy 1.0 and 1.0 for Iris, 0.95 and 1.0 for Spiral, 0.835 and 0.71 for Telecom, respectively), because our properties and metrics are ML model performance-agnostic.

We use the following explainers for obtaining explanations: SHAP (Lundberg and Lee 2017) shap 0.46.0, Lime (Ribeiro et al. 2016a) lime 0.2.0.1 and Integrated Gradients (IG) (Sundararajan et al. 2017) via captum 0.7.0 (Kokhlikyan et al. 2020) for feature importance-based explanations with all the models (except for IG with XGB, because IG work with gradients, whereas XGB tree ensembles are non-differentiable); Baseline (Wachter et al. 2018) counterfactuals via mlxtend 0.23.4 (Raschka 2018) and DiCE (Mothilal et al. 2020) dice-ml 0.11 counterfactuals for counterfactual explanations with all the models (except for DiCE with binary classifiers, thus omitting DiCE with models trained on Iris, since the off-the-shelf cross-compatible version of dice-ml works only for binary classification); Anchors (Ribeiro et al. 2016b) via alibi 0.9.6 (Klaise et al. 2021) for rule-based explanations with all the models. We chose these explainers because we found them to be the only ones immediately available (via the Python Package Index (PyPi)), easily usable with XGB and NN models trained on datasets not requiring

specific pre-processing, as well as cross-compatible among other libraries without modifications. This way our results can be reproduced using only the specifications (aside from the proprietary Telecom dataset).

We will specify the relevant explainer parameters when discussing specific explanations in the next section. We will likewise specify the functions used for implementing the specific instantiations of metrics for quantifying explanation correctness.

We reiterate that our empirical setup is mostly for illustration purposes. We aim to provide a feeling as to measuring explanation correctness using our framework, but do not claim significant experimental findings. Extensive empirical evaluation of various explainers with different ML classifiers is prohibitively costly for a research paper, since particularly counterfactual and rule-based explanation generation is extremely time-consuming (even our final experiments took 15 hours on a powerful MacBook.) We welcome practitioners to use our framework to assess correctness of explanations from particular explainers (varying their hyperparameters) applied to concrete models trained on relevant datasets for a downstream task at hand. Our main contributions in this paper are instead conceptual and theoretical.

In the next section, we study whether we can formulate desirable properties and formal metrics for assessing correctness of three forms of explanations (feature importance-based, rule-based and counterfactual) of ML model input/output behaviour with respect to classification tasks on tabular data. We present two such properties, namely soundness and completeness, and state multiple metrics for measuring how well the different forms of explanations satisfy the two properties.

4 Correctness of Importance-Based, Rule-Based and Counterfactual Explanations of ML Model Input/Output Behaviour

Correctness pertains to faithfulness or truthfulness of explanations with respect to the underlying ML model (Nauta et al. 2023, p. 10). Colloquially, we say that an explanation is correct if the explanation is *sound* in the sense that the model behaves the way the explanation says, and if the explanation is *complete* in the sense that it covers the model's behaviour. In this, we try to adhere to the notions of soundness and completeness from formal logics, and follow very closely the early work of (Kulesza et al. 2013) that introduced these notions analogously.[5] We thus submit two distinct properties that address correctness of explanations, namely soundness and completeness, on which we elaborate next.

The property of *soundness* addresses how truthful an explanation is with respect to the underlying ML model. In simple terms, soundness requires that if there is an explanation, then the model behaves the way the explanation describes. Or as originally albeit informally put (Kulesza et al. 2013, p. 4), soundness captures "nothing but the truth", namely "the extent to which each component of an explanation's content is truthful in describing the underlying system."

The property of *completeness* addresses how well an explanation generalises, so as to cover the underlying ML model well-enough. In simple terms, completeness requires that for any model behaviour there is an explanation. As originally put (Kulesza et al. 2013, p. 4), completeness captures "the whole truth", namely "the extent to which all of the underlying system is described by the explanation."

We note that soundness and completeness of explanations are sometimes orthogonal in the sense that achieving one does not necessarily help achieving the other. As noted in (Sokol and Flach 2020), model-agnostic explainers may be unable to take advantages of the ML model internals and thus entail a trade-off in terms of soundness and completeness of explanations. For example, a rule generated by sampling model input/output instances may be highly sound in that the underlying ML model predictions for the instances covered by the rule would match the rule conclusion – in the well-known Anchors approach (Ribeiro et al. 2018) terminology, the rule is at least 95%

[5] In (Kulesza et al. 2013), and likewise in the work (Sokol and Flach 2020), the various authors use the word 'fidelity' for what we term correctness, but propose soundness and completeness as the two dimensions/properties/metrics for assessing fidelity/correctness.

'precise'. However, the rule may not be very complete in that only a small part of the dataset would be covered by the rule – 2% 'coverage' as in (Ribeiro et al. 2018). Conversely, such a rule may be highly complete in covering a large region of the dataset, but not entirely sound in matching the predictions of the underlying ML model for some (unsampled) instances in that region.

More concretely regarding correctness of explanations, we next consider what soundness and completeness mean for the three different forms of explanations, namely rule-based, counterfactual and feature importance-based ones. In what follows, we discuss one by one each of the forms of explanations, stipulating first soundness and then completeness properties, presenting the relevant metrics alongside. We will speculate as to whether satisfying one or another property is desirable and how this reflects on the metrics for assessing each of the properties. We give a bird's eye view of the schematic connections in Figure 1, with concrete properties and their metrics linked to their definitions.

4.1 Correctness of Rule-Based Explanations

Intuitively, rule-based explanations (see Definition 3.1) are correct if each of them truthfully classifies instances it covers (sound) and they collectively cover all inputs (complete). A formal notion of cover will be crucial to both soundness and completeness of rule-based explanations. To this end, we first introduce an auxiliary definition of what it means for an instance to be covered by a rule, namely that the instance's feature values fall within the (sets of ranges of) values specified in the rule's antecedent.

Definition 4.1. A rule-based explanation $R = (\{f_1 : s_1, \ldots, f_n : s_n\}, c)$ **covers** an input instance $\mathbf{x} = (x_1, \ldots, x_{|\mathbf{X}|}) \in \prod_i^{|\mathbf{X}|} D_i$ just in case $x_k \in s_k \; \forall f_k \in \{f_1, \ldots, f_n\}$. We also say that $\mathbf{x}$ is *covered* by R if R covers $\mathbf{x}$.

The **cover**[6] of a set $E = \{R_1, \ldots, R_m\}$ of rule-based explanations is defined as $cover(E) \coloneqq \{\mathbf{x} \in X \; : \; \text{some } R_i \in E \text{ covers } \mathbf{x}\}$. (We can also treat a single explanation R as a singleton set $E = \{R\}$ for the purposes of finding its cover $cover(R) = cover(\{R\})$.)

We are now in position to formulate the soundness property for rule-based explanations.

4.1.1 *Soundness of Rule-based Explanations.*

We essentially want to say that a rule-based explanation is sound to the degree that the rule's conclusion matches the ML model outputs for the data instances covered by the rule.

Property 1 (Soundness of rule-based explanations). *A (local or global) rule-based explanation $E = (\{f_1 : s_1, \ldots, f_n : s_n\}, c)$ is **sound** in as much as its conclusion agrees with the ML model output, i.e. $\mathcal{M}(\mathbf{x}') = c$, for instances $\mathbf{x}'$ covered by E.*

We posit that it is evidently desirable for rule-based explanations to be as sound as possible. We thus turn to consider how to measure soundness of rule-based explanations.

Property 1 seems straightforward in that it requires rule-based explanation conclusions to match model outputs for rule-covered instances. A metric for this is typically known as *fidelity* (see e.g. (Kulesza et al. 2013)), which measures "the fraction of data samples for which predictive model [$\mathcal{M}$] and an explanation make the same decision" (Nauta et al. 2023, p. 22). Given our notions of rule-based explanations and cover (see Definitions 3.1 and 4.1), we measure fidelity of the former in terms of ratio of instances covered by the explanation and matching the model output against all the covered instances.

Fidelity of Rule-based Explanations.

[6]The definition of rule cover is adopted from (Lakkaraju et al. 2019).

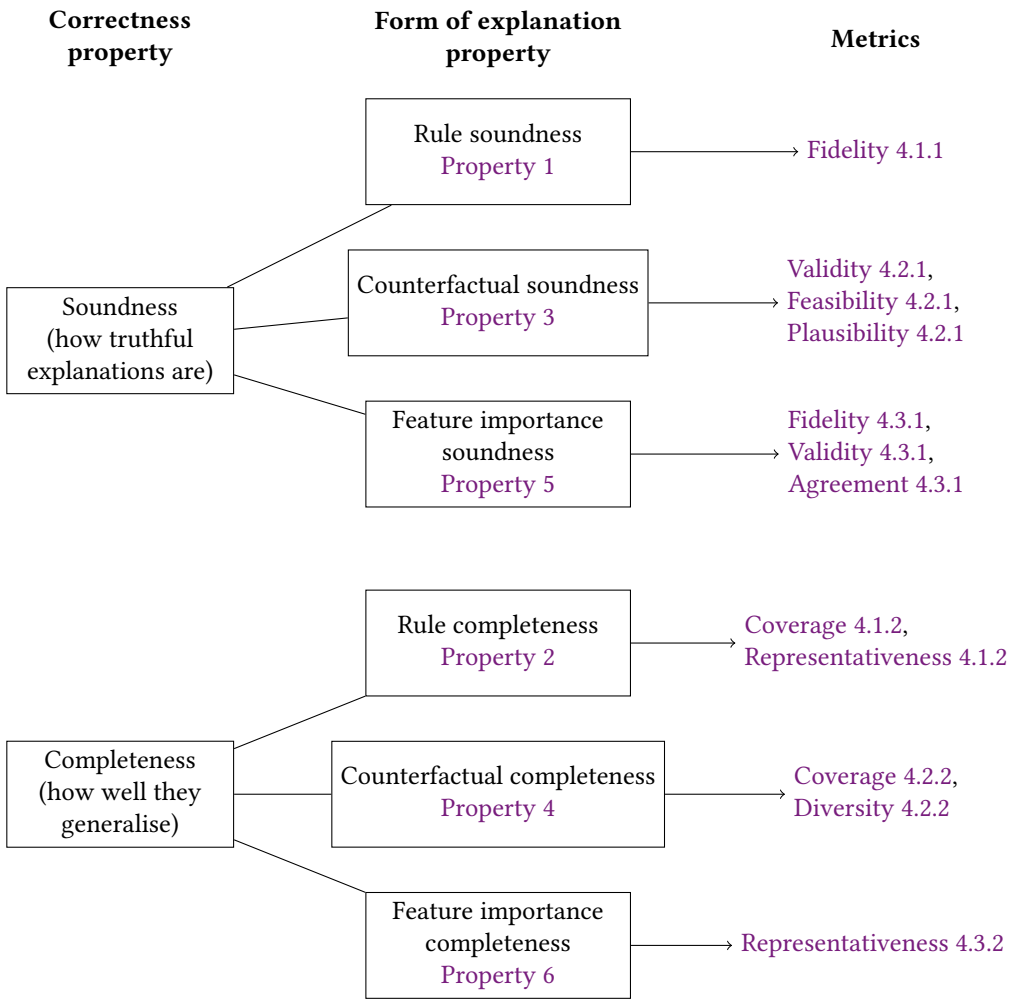

Fig. 1. A schematic view of properties and metrics for assessing correctness of three forms of explanations in terms of their soundness (colloquially, how truthful explanations are) and completeness (colloquially, how well the explanations generalise). The formal property and metric definitions are numbered and linked.

Metric 1. *The **fidelity** metric $Q^{FIDELITY}$ for measuring soundness of a (local or global) rule-based explanation $R = (\{f_1 : s_1, \ldots, f_n : s_n\}, c)$ or a set $E = \{R_1, \ldots, R_m\}$ of explanations assigns to $\{R\}$ or E its* fidelity score *thus:*[7]

$$Q^{FIDELITY}(E) = \frac{|\{\mathbf{x}' \in cover(E) \ : \ \mathcal{M}(\mathbf{x}') = c\}|}{|cover(E)|}. \tag{3}$$

[7]In (Lakkaraju et al. 2019), the numerator from our equation measures rule 'disagreement' score as part of quantifying 'fidelity' of rule-based explanations.

Colloquially, a rule-based explanation is entirely faithful if $\mathcal{M}(\mathbf{x}') = c$ for all $\mathbf{x}' = (x'_1, \ldots, x'_{|\mathbf{X}|})$ covered by E.[8] We maintain it is highly desirable to have entirely faithful rule-based explanations – after all, if an explanation says that under the premises concerning input feature values the output will be so and so, then that should be a true statement if it is to explain the model behaviour.

Example 4.2. Revisiting Example 3.2, consider the rule-based explanations $\mathcal{R} = \{R_1 = (\{f_1 : 1\}, 1), R_2 = (\{f_2 : 1\}, 1), R_0 = (\{f_1 : 0, f_2 : 0\}, 0)\}$ for the model $\mathcal{M}$ representing the logical OR function of two binary variables: $\mathcal{M}(x_1, x_2) = x_1 \vee x_2$ for $(x_1, x_2) \in \{0, 1\}^2$. Explanation R_1 covers instances $(1, 0)$ and $(1, 1)$, so that $cover(R_1) = \{(1, 0), (1, 1)\}$. Clearly, on both inputs $\mathcal{M}$ outputs 1, so that $Q^{\text{FIDELITY}}(R_1) = 1$, i.e. R_1 is an entirely faithful explanation. A similar argument shows that both R_0 and R_2 are entirely faithful too.

An unfaithful explanation could for example be $R' = (\{f_1 : 1, f_2 : 0\}, 0)$, which says that the instance $(1, 0)$ should be classified as 0. This rule has fidelity score 0, since the only instance it covers is classified differently by $\mathcal{M}$ than R' says. Clearly, such a rule does not speak truth about $\mathcal{M}$.

As part of our empirical setup described in Section 3.4, we use the Anchors explainer alibi.explainers.AnchorTabular, fitting on the train datasets X_{train}, with numerical features discretised into 10 evenly spaced bins (percentiles disc_perc=[10, 20, …, 90]). We generate Anchor rule-based explanations for all instances of each of the three test datasets X_{test} with corresponding XGB and NN models, using three values of the 'minimum precision' parameter threshold, namely 1.0, 0.95 (default) and 0.8. We report the results in Table 1: we report the fidelity of the set of Anchor explanations generated for the instances from the test set with respect to instances from the entire dataset (i.e. train and test) covered by any of the explanations.

Table 1. Fidelity scores (from now on, rounded to three decimal places) of Anchor rule-based explanations generated with respect to XGB and NN models for instances from test datasets X_{test} (their sizes, equal to the numbers of explanations, indicated in brackets), varying precision threshold values, with rule cover taken over the entire datasets X.

Model type	**Dataset** ($\lvert X_{test}\rvert$ = #**expls**)	threshold=1.0	threshold=0.95	threshold=0.8
XGB	Iris (35)	0.980	0.980	1.000
	Spiral (20)	0.957	0.957	0.938
	Telecom (200)	0.981	0.988	1.000
NN	Iris (35)	1.000	0.986	1.000
	Spiral (20)	0.987	0.987	0.936
	Telecom (200)	1.000	1.000	1.000

Note that even with threshold = 1, not all Anchors explanations are entirely faithful, at least for some models and datasets. Some examples:

- Iris instance 84 with feature values $(1.085, -0.137, 0.729, 0.689)$, here and henceforth rounded to three decimal places) is classified by the XGB model in class 1, but is covered by the Anchor rule-based explanation $(\{1 : (-\infty, \infty), 2 : (-\infty, \infty), 3 : (0.689, \infty), 4 : (-\infty, \infty)\}, 2)$ that has class 2 as a consequent;
- Spiral instance 26 with feature values $(-1.675, -0.928)$ is classified by the NN model in class 1, but is covered by rule $(\{1 : (-\infty, -0.746], 2 : (-0.938, \infty)\}, 2)$ with consequent 2;
- Telecom instance 301 with $(-1.076, -1.069, -1.081, -1.042, -1.058, 0.016, -0.559, 0.257, -0.942, -0.938, -0.943, -1.089)$ is classified as 0, but is covered by $(\{1 : (-\infty, -0.667],$

[8] In (Ignatiev 2020), this is referred to as a 'correct' rule-based explanation.

$2 : (-\infty, \infty), 3 : (-\infty, \infty), 4 : (-\infty, \infty), 5 : (-1.303, \infty), 6 : (-\infty, 0.203], 7 : (-\infty, -0.295], 8 : (-\infty, \infty), 9 : (-1.714, \infty), 10 : (-1.682, \infty), 11 : (-1.661, \infty), 12 : (-1.194, \infty)\}, 1)$.

We also see that higher precision threshold values do not in general correlate with higher fidelity: for example, explanations for the XGB model on Telecom are entirely faithful with threshold 0.8 but have fidelity score 0.981 with threshold 1. This is because fidelity is an analytically defined metric with exact value for any explanation, whereas the Anchor precision threshold plays role in a process of sampling instances and rules when generating explanations. But overall, Anchors rule-based explanations exhibit high fidelity, as desired.

We now consider the completeness property for rule-based explanations.

4.1.2 Completeness of Rule-based Explanations. Recall from the beginning of Section 4 that the property of explanation completeness addresses how well the explanations cover the ML model's input/output behaviour. With respect to rule-based explanations, we essentially want to say they are complete in as much as that they can explain the model's output for any input, i.e. cover all possible inputs, and are representative, or exhaustive, of the model's behaviour, i.e. cover all the ways of obtaining the model's output.

Property 2 (Completeness of rule-based explanations). *A rule-based explanation $E = (\{f_1 : s_1, \ldots, f_n : s_n\}, c)$ is **complete** in as much as it covers the underlying ML model's input/output pairs and represents the rules that explain the model.*

Note that we could talk about completeness of a set of explanations, instead of a singular explanation, in that typically we expect multiple rules to represent the whole underlying ML model and cover any possible output. The two ways can however be seen as equivalent. Indeed, a rule-based explanation $R_1 = (\{f_1 : s_1, \ldots, f_n : s_n\}, c)$ can equivalently be expressed (with an abuse of notation) as a logical implication $f_1 : s_1 \wedge \ldots \wedge f_n : s_n \rightarrow c$, and a finite collection $R_1, \ldots, R_m$ of such rules can be expressed as a single disjunction $R = R_1 \vee \ldots \vee R_m$. Then either R or $R_1, \ldots, R_m$ can be equivalently inspected for coverage and representation of $\mathcal{M}$. We may choose one or the other formulation depending on which sounds more natural.

We posit that it is desirable for rule-based explanations to be as complete as possible. Next, we consider ways to measure completeness of rule-based explanations.

One key idea behind measuring completeness of rule-based explanations is that of rule's *coverage* in terms of instances covered by the explanation, see e.g. (Ignatiev 2020; Lakkaraju et al. 2019). For a single rule-based explanation this can be measured in terms of the covered instances for the output class of the rule. For a set of rule-based explanations it can instead be measured in terms of all the covered instances. We state the coverage metric as follows.

Coverage of Rule-based Explanations.

Metric 2. *The **coverage** metric $Q^{COVERAGE}$ for measuring completeness of a rule-based explanation $R = (\{f_1 : s_1, \ldots, f_n : s_n\}, c)$ or a set $E = \{R_1, \ldots, R_m\}$ of explanations assigns to R or E, respectively, its* coverage value *(see Definition 4.1) thus:*

$$Q^{COVERAGE}(R) = \frac{|cover(\{R\})|}{|\{\mathbf{x}' \in X \ : \ \mathcal{M}(\mathbf{x}') = c\}|}; \tag{4}$$

$$Q^{COVERAGE}(E) = \frac{|cover(E)|}{|X|}. \tag{5}$$

Typically, the higher the coverage of a rule-based explanation, the better. Note though that while the coverage value of a single local rule-based explanation need not be high, multiple of those could well cover the whole class. Similarly, a single global rule-based explanation need not be expected to obtain coverage value 1, but one could treat several such explanations as a single rule (as per the note immediately after Property 2) to cover the

whole class, and ultimately the whole dataset. Ideally then, a set of global rule-based explanations would have full coverage of the dataset X with coverage value 1, so that every input is covered by (at least) one explanation.

Example 4.3. Recall the explanations $\mathcal{R} = \{R_1 = (\{f_1 : 1\}, 1), R_2 = (\{f_2 : 1\}, 1), R_0 = (\{f_1 : 0, f_2 : 0\}, 0)\}$ from Example 4.2 for our toy model $\mathcal{M}$ that learnt the logical OR function of two binary variables: $\mathcal{M}(x_1, x_2) = x_1 \vee x_2$ for $(x_1, x_2) \in \{0, 1\}^2$. Example 4.2 shows that $cover(R_1) = \{(1, 0), (1, 1)\}$. Since there are 3 instances classified as 1 by $\mathcal{M}$, we find $Q^{\text{COVERAGE}}(R_1) = 2/3$. The same holds for $Q^{\text{COVERAGE}}(R_2)$.

Now while R_1 and R_2 together cover the output class 1, they do not cover all of the inputs. In particular, $(0, 0)$ is covered only by R_0. So $Q^{\text{COVERAGE}}(\mathcal{R}) = 1$. Note also that $Q^{\text{COVERAGE}}(\{R_0, R_1\}) = 3/4$, and likewise for $\{R_0, R_2\}$, so that $\mathcal{R}$ is the only set of rule-based explanations that fully covers $\mathcal{M}$.

Considering our empirical setup, similarly to Table 1, we report in Table 2 the coverage of the set of Anchors explanations generated for the instances from the test sets. Note that not all sets of explanations (of test instances) fully cover the model, depending on the model and the dataset it is trained on. But overall, Anchors rule-based explanations exhibit high coverage, as desired.

Table 2. Coverage values of Anchor rule-based explanations generated with respect to XGB and NN models for instances from test datasets with varying precision threshold values.

Model type	**Dataset** ($\lvert X_{test}\rvert$ = #**expls**)	threshold=1.0	threshold=0.95	threshold=0.8
XGB	Iris (35)	0.993	0.993	1.000
	Spiral (20)	0.920	0.920	0.970
	Telecom (200)	1.000	1.000	1.000
NN	Iris (35)	0.960	0.980	0.987
	Spiral (20)	0.790	0.790	0.940
	Telecom (200)	1.000	1.000	1.000

Another key idea related to coverage is that of *representativeness*, which in simple terms aims to measure whether a set of rule-based explanations represents all the rules that may explain the model's output. We are not aware of this idea formalised in the XAI literature and will thus next give what we believe to be a novel metric for measuring completeness of rule-based explanations.[9] Let us explain and formalise the idea.

First, to judge whether a given set E of rule-based explanations represents all the relevant kind of rule-based explanations that can explain the model's outputs, we need to define the latter collection. In other words, we want to define the space of all the rules. So let us define $\mathcal{R}$ to be the set of all rule-based explanations obtainable from the explainer EX.

Next, we want to focus on the possible rules that also cover the inputs that are covered by E. We know the inputs covered by E, it is $cover(E)$. So we can ask, what are the rules that also cover the inputs covered by E? To this end, we define $\mathcal{A}_E \coloneqq \{R \in \mathcal{R} \ : \ R \text{ covers some } \mathbf{x} \in cover(E)\}$. This is the set of possible rules that cover at least something covered by E. In other words, $\mathcal{A}_E$ comprises all the rule-based explanations that could explain input/output pairs explained by E. Intuitively, $\mathcal{A}_E$ captures maximal representation of E.

So we ask, does $\mathcal{A}_E$ cover any more inputs than E does? If the answer is no, i.e. $E = \mathcal{A}_E$, then E represents all the explanations of input/output pairs explained by E. We would in such case say that E is entirely representative.

[9]We note, however, that the metric we call 'coverage' is called 'representativeness' in (Carvalho et al. 2019); and it also corresponds to two concepts called 'representativeness' and 'accuracy' in (Robnik-Šikonja and Bohanec 2018). Those notions of representativeness are different from ours.

If, on the other hand, the answer is yes, i.e. $cover(E) \subsetneq cover(\mathcal{A}_E)$, then E is not entirely representative of the rules that could explain what it explains. We state the representativeness metric formally before giving an example.

Representativeness of Rule-based Explanations.

Metric 3. *The* ***representativeness*** *metric Q^{REPR} for measuring completeness of a set $E = \{R_1, \ldots, R_m\}$ of rule-based explanations assigns E its* representation value *as the ratio of the number of instances covered by (rules in) E against the number of instances that would be covered by any rule-based explanation that could cover something covered by E thus:*

$$Q^{REPR}(E) = \frac{|cover(E)|}{|cover(\mathcal{A}_E)|} = \frac{|cover(E)|}{|cover(\{R \in \mathcal{R} \ : \ R \text{ covers some } \mathbf{x} \in cover(E)\})|}. \tag{6}$$

Note that we can treat a single explanation $E_{\mathbf{x}}$ as a singleton $\{E_{\mathbf{x}}\}$ for the purposes of finding its representation value.

Example 4.4. Continuing Example 4.3, as R_2 covers $(1,1) \in \{(0,1),(1,1)\} = cover(\{R_1\})$, we have $R_2 \in \mathcal{A}_{\{R_1\}}$. Indeed, the cover of $\mathcal{A}_{\{R_1\}}$ properly contains the cover of $\{R_1\}$: we have $cover(\{R_1\}) \subsetneq cover(\mathcal{A}_{\{R_1\}}) = cover(\mathcal{R} \setminus \{R_0\}) = \{(0,1),(1,0),(1,1)\}$. In particular, $Q^{\text{REPR}}(\{R_1\}) = \frac{|cover(\{R_1\})|}{|cover(\mathcal{A}_{\{R_1\}})|} = \frac{|\{(0,1),(1,1)\}|}{|\{(0,1),(1,0),(1,1)\}|} = 2/3$. An analogous argument shows $Q^{\text{REPR}}(\{R_2\}) = 2/3$ too.

On the other hand, observe that the explanation set $E_{12} = \{R_1, R_2\}$ consisting of both minimal explanations for model output 1 satisfies $E_{12} = \mathcal{A}_{E_{12}}$, and hence is entirely representative with representation value 1. Yet recall from Example 4.3 that E_{12} does not fully cover the model input space. Instead, only $\mathcal{R}$ itself does. And $\mathcal{R}$ is also (trivially) entirely representative: $Q^{\text{REPR}}(\mathcal{R}) = 1$.

Similarly to Table 2, we report in Table 3 the representativeness of the set of Anchors explanations generated for the instances from the test sets. Note that not all sets of explanations (of test instances) entirely represent all the relevant explanations, depending on the model and the dataset it is trained on. But overall, Anchors rule-based explanations exhibit high representativeness, as desired.

Table 3. Representation value of Anchor rule-based explanations generated with respect to XGB and NN models for instances from test datasets with varying precision threshold values.

Model type	**Dataset** ($\lvert X_{test}\rvert$ = #**expls**)	threshold=1.0	threshold=0.95	threshold=0.8
XGB	Iris (35)	0.993	0.993	1.000
	Spiral (20)	0.920	0.920	0.970
	Telecom (200)	1.000	1.000	1.000
NN	Iris (35)	0.960	0.980	0.987
	Spiral (20)	0.790	0.790	0.940
	Telecom (200)	1.000	1.000	1.000

To sump up, for a set of rule-based explanations to be complete we are looking for it to be fully covering (of the model inputs) and entirely representative (of the rules that explain the model's input/output behaviour). After all, a complete set of explanations should explain all the ways the model outputs something, for any input.

We now turn to counterfactual explanations.

4.2 Correctness of Counterfactual Explanations

The fundamental idea behind counterfactual explanations (see Definition 3.1) is that they identify situations where the model output diverges while the input stays as similar as possible (Wachter et al. 2018, p. 9). This suggests a straightforward if rather broad formulation of counterfactual soundness.

4.2.1 Soundness of Counterfactual Explanations. We deem a counterfactual explanation sound if it describes a possible situation which is classified differently than the original one.

Property 3 (Soundness of counterfactual explanations). *A counterfactual explanation $E_{\mathbf{x}}$ is **sound** in as much as the counterfactual is possible and the ML model's outputs for $E_{\mathbf{x}}$ and the explained instance $\mathbf{x}$ differ.*

The word 'possible' is intentional here, as we realistically do not want counterfactuals that are beyond the possible worlds/situations given the input instance (see (Ginsberg 1986) for an exposition to counterfactuals). We posit that it is evidently desirable for counterfactual explanations to be sound. We now consider how to measure soundness of counterfactual explanations.

With Property 3 in mind, it seems straightforward to say if and when a counterfactual explanation is sound, namely when the changes to feature values to the original instance result in changed model output. This aspect is known as *validity* (see e.g. (Guidotti 2022)) and is easy to quantify. Stated simply, counterfactual is valid if it changes the model output from the original one to the desired one, provided that the latter two are different. It is often that an explainer yields a set of counterfactual explanations for a given instance, in which case the fraction of valid counterfactual explanations measures the validity of the set. We thus state the validity metric for single explanations and sets thereof.

Validity of Counterfactual Explanations.

Metric 4. *The **validity** metric Q^{VALIDITY} for measuring soundness of a counterfactual explanation $E_{\mathbf{x}} = \mathbf{x}' = (f_1 : v_1, \ldots, f_{|\mathbf{X}|} : v_{|\mathbf{X}|})$ or a set $E = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_m\}$ of counterfactual explanations assigns to $E_{\mathbf{x}}$ or E, respectively, its* validity score *thus:*

$$Q^{\text{VALIDITY}}(E_{\mathbf{x}}) = \begin{cases} 1 \text{ if } \mathcal{M}(\mathbf{x}') \neq \mathcal{M}(\mathbf{x}), \\ 0 \text{ otherwise;} \end{cases} \tag{7}$$

$$Q^{\text{VALIDITY}}(E) = \frac{|\{\mathbf{x}'_i \in E \ : \ \mathcal{M}(\mathbf{x}'_i) \neq \mathcal{M}(\mathbf{x})\}|}{m}. \tag{8}$$

Colloquially, a counterfactual is (entirely) valid with validity score 1 if it actually changes the classification. A set of counterfactuals has validity score directly proportional to the number of valid counterfactuals it contains.

Example 4.5. Revisiting Example 3.2, consider the counterfactual explanations $C = \{\mathbf{x}'_1 = (f_1 : 1, f_2 : 0), \mathbf{x}'_2 = (f_1 : 0, f_2 : 1), \mathbf{x}'_3 = (f_1 : 1, f_2 : 1)\}$ for input $\mathbf{x} = (0, 0)$ to model $\mathcal{M}$. They are clearly all (entirely) valid, since $\mathcal{M}$ models logical disjunction so that setting any feature value to 1 results in input $\mathbf{x}' \in C$ with output $\mathcal{M}(\mathbf{x}') = 1 \neq 0 = \mathcal{M}(\mathbf{x})$.

On the other hand, for input $(1, 1)$ the counterfactual $\mathbf{x}' = (f_1 : 0)$ would be invalid with validity score 0, because $\mathcal{M}(\mathbf{x}') = \mathcal{M}(0, 1) = 1 = \mathcal{M}(1, 1)$.

In multi-class classification problems, one could also specify the desired class of a counterfactual. In that case, the above validity metric can be easily modified to take the desired class label c into account, counting a counterfactual $\mathbf{x}'$ valid only if $\mathcal{M}(\mathbf{x}') = c$, and accordingly for a set of counterfactuals.

As part of our empirical setup described in Section 3.4, we use what we call the Baseline explainer via the mlxtend.evaluate.create_counterfactual method that implements counterfactual generation as described by (Wachter et al. 2018). We set the parameter y_desired_proba controlling how confident the model is when classifying

a potential counterfactual to 1.0, aiming to ensure higher validity of counterfactuals. Baseline counterfactual explanations also depend on the regularisation parameter lammbda=λ [sic], where $\lambda \in [0, \inf)$ controls how significant it is for the predicted class probability of the counterfactual to be different from that of the input instance. That is, when increased, λ encourages bigger penalty for higher squared difference between the model predicted probabilities for the input instance and its counterfactual. We also use the DiCE explainer, loading XGB models with backend='sklearn' and NN models with backend='PYT', in both cases using method='random' for generating counterfactuals. Other DiCE parameters are default.

We generate Baseline and DiCE counterfactual explanations $E = \{\mathsf{EX}(\mathcal{M}(\mathbf{x}) = y) \ : \ \mathbf{x} \in X_{test}\}$ for test instances of each of the three datasets X with corresponding XGB and NN models $\mathcal{M}$ (except from the XGB model trained on Iris, because the cross-compatible version of the DiCE explainer that we use works for binary classification only). We report the validity of Baseline and DiCE counterfactual explanations in Table 4.

Table 4. Validity scores of DiCE and Baseline counterfactual explanations generated with respect to XGB and NN models for instances from test datasets, with varying Baseline regularisation parameter λ values.

Model type	**Dataset** ($\lvert X_{test}\rvert$ = #**expls**)	DiCE	Baseline				
			$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 10$	$\lambda = 100$
XGB	Iris (35)	–	0.3	0.5	0.6	0.6	0.633
	Spiral (20)	1.0	0.0	0.5	0.65	0.7	0.7
	Telecom (200)	1.0	0.44	0.48	0.48	0.48	0.465
NN	Iris (35)	–	0.567	0.567	0.633	0.667	0.7
	Spiral (20)	1.0	0.0	0.25	0.45	0.5	0.8
	Telecom (200)	0.99	0.17	0.17	0.175	0.2	0.175

As desired, DiCE counterfactuals are almost always entirely valid. This is ensured by the explanation method, and the validity score only happens to be 0 when no counterfactual is generated, which sometimes happen due to the underlying search procedure terminating before finding one. Baseline counterfactuals, on the other hand, exhibit low validity scores. With increasing the λ parameter that effectively encourages smaller distance between the input instance and its counterfactual, Baseline counterfactual validity typically increases for models trained on Iris and Spiral; though for the two models trained on Telecom, explanation validity seems to peak with $\lambda = 10$.

Now, while validity measures whether or not the ML model outputs for the original instance and its counterfactual differ, it does not say anything about the possibility of the changes in the input. Indeed, with counterfactuals it is imperative to remember that not all features are created equal, in the sense that some feature values cannot be altered altogether or can only change to certain values. For example, in a variation of the by now classic example of a bank loan application (see e.g. (Guidotti 2022)), the decision may be based on a person's age, and if demanded a counterfactual explanation as to why a loan is not granted, it would not be very useful to produce a counterfactual in which the person gets younger. (It is not in principle absurd to have such an explanation, for it may be interesting to the person to know that had they been younger, things would have been different; but in the spirit of counterfactuals as closest possible worlds (Ginsberg 1986), we want to distinguish between actually possible situations and those that are possible only in principle.) In other words, we are talking about *constraints*

that cannot be violated when generating counterfactuals. This aspect is known as *feasibility* (see e.g. (Guidotti 2022; A.-H. Karimi et al. 2020))[10] and we next consider how to measure it.

The in-principle *feasibility* of counterfactuals is measured by identifying any constraints violated through generating an explanation. We first define what we mean by a constraint on a counterfactual and then show how to measure constraint satisfaction.

Definition 4.6. A (counterfactual) **constraint** is a pair $C = (f_i : D_i^*)$ of feature $f_i \in \mathbf{X}$ and a set $D_i^* \subseteq D_i$ of values in its domain.
A counterfactual explanation $E_\mathbf{x} = (f_1 : v_1, \ldots, f_{|\mathbf{X}|} : v_{|\mathbf{X}|})$ *satisfies* constraint $C = (f_i : D_i^*)$ just in case $v_i \in D_i^*$.

If a counterfactual explanation does not satisfy a given constraint, we may say it *violates* the constraint.

Intuitively, a constraint on a counterfactual explanation is simply a restriction of some feature's values to some domain. Obviously, multiple constraints can be imposed on a counterfactual at once. Measuring satisfaction of possibly multiple constraints is what feasibility amounts to.

Feasibility of Counterfactual Explanations.

Metric 5. *The* ***feasibility*** *metric $Q^{\text{FEASIBILITY}}$ for measuring soundness of a counterfactual explanation $E_\mathbf{x} = (f_1 : v_1, \ldots, f_{|\mathbf{X}|} : v_{|\mathbf{X}|})$ given a set $\mathbf{C} = \{C_1, \ldots, C_k\}$ of constraints assigns to $E_\mathbf{x}$ its* feasibility score *thus:*

$$Q^{\text{FEASIBILITY}}(E_\mathbf{x}) = \frac{|\{C_i \in \mathbf{C} \ : \ E_\mathbf{x} \text{ satisfies } C_i\}|}{k}. \tag{9}$$

Colloquially, a counterfactual explanation is entirely feasible with feasibility score 1 if its feature values are constrained as desired.

Example 4.7 (Example 4.5 continued.). Suppose that there was a constraint $C = (f_1 : \{1\})$ on counterfactual explanations in Example 3.2. That is, any counterfactual was constrained to have f_1 value 1. In C, only $\mathbf{x}'_1$ and $\mathbf{x}'_3$ satisfy C, whereas $\mathbf{x}'_2$ violates it. The former two are thus entirely feasible, whereas $Q^{\text{FEASIBILITY}}(\mathbf{x}'_2) = 0$.

Let us illustrate with the DiCE and Baseline explanations. We completely constrain DiCE to be able to vary the values of only the following features (with respect to the three datasets): Iris features 2 and 4; Spiral feature 1; Telecom features 6, 8. For example, for any Spiral test instance $\mathbf{x} = (v_1, v_2)$, the constraint formally is $C = (f_2 : v_2)$, with either the XGB or NN model, meaning that to satisfy C the counterfactual $E_\mathbf{x} = (f_1 : v'_1, f_2 : v'_2)$ must have $v'_2 = v_2$ and any v'_1.

The choices of constrained features for Iris and Spiral are ad-hoc, whereas those for Telecom come from expert knowledge. Notably, since DiCE strictly enforces any constraints given, it may fail to generate counterfactuals in the presence of constraints if the underlying search procedure does not find one satisfying the constraints. In such a case, we may deem the feasibility score of the (non-existent) counterfactual explanation $\emptyset$ to be either 0 or 1. As a consequence, if we deem it 0, then the average feasibility score of DiCE counterfactual explanations need not be 1, even though DiCE respects constraints. If we deem it 1, then DiCE average feasibility should be 1. Meanwhile, the Baseline explainer does not allow for a simple way to input constraints (other than modifying the algorithms), and in general does not satisfy them. We report the *average* feasibility of DiCE and Baseline counterfactual explanations in Table 5: we average the individual feasibility scores across the explanations.

First note that the feasibility of Baseline counterfactual explanations is essentially zero (irrespective of λ), except for some counterfactuals for Iris instances that by chance happen to satisfy the constraints. Meanwhile, DiCE counterfactual explanations exhibit highly varied feasibility scores on average, because while any particular generated explanation has fidelity score 1 by construction, for so many input instances no counterfactual is

[10]Though it is sometimes referred to as 'plausibility', e.g. in (A. Karimi et al. 2023), we reserve the latter term for another metric (similarly to (Guidotti 2022)) to be discussed in turn.

Table 5. Feasibility scores of DiCE and Baseline counterfactual explanations generated with respect to XGB and NN models for instances from test datasets. For DiCE, when no counterfactual is generated due to constraints, we choose to assign the feasibility score of either 0 ($\emptyset = 0$) or 1 ($\emptyset = 1$). For Baseline, we vary the regularisation parameter λ to indicate that it makes no difference to satisfaction of constraints.

Model type	**Dataset** ($\lvert X_{test}\rvert$ = #**expls**)	**Constrained feature indices**	DiCE		Baseline		
			$\emptyset = 0$	$\emptyset = 1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 10$
XGB	Iris (35)	1, 3	–	–	0.033	0.033	0.033
	Spiral (20)	2	1.0	1.0	0.0	0.0	0.0
	Telecom (200)	$\{1, \ldots, 12\} \setminus \{6, 8\}$	0.03	1.0	0.0	0.0	0.0
NN	Iris (35)	1, 3	–	–	0.033	0.033	0.033
	Spiral (20)	2	0.35	1.0	0.0	0.0	0.0
	Telecom (200)	$\{1, \ldots, 12\} \setminus \{6, 8\}$	0.26	1.0	0.0	0.0	0.0

generated. By agreement, this yields individual fidelity scores of either 0 or 1, which moves the average feasibility either closer to 0 or all the way to 1.

We note that there is a related notion of 'actionability', see e.g. (Sokol and Flach 2020).[11] It refers to the explanation providing the user guidance as to how to affect the model's decision (i.e. output), typically towards a desired outcome. In the bank loan example, a counterfactual explanation that refers to, say, reducing the number of one's active loans is more actionable than one suggesting to get younger. So in principle it sounds like feasibility, but requires some sort of choice as to what is actionable from the user's perspective. It may be instructive to think about feasibility measuring adherence to *hard* constraints, in the parlance of optimisation theory. The sister notion of *soft* constraints, i.e. those that can be violated albeit at a penalty, is analogous to constraints whose satisfaction could be measured by some actionability metric. We do not consider the latter here, since we focus on functionally grounded rather than user-dependent metrics.

Constraint violation is not the only option for measuring how possible counterfactuals are. Another aspect is that of *plausibility*, which accounts for how close the counterfactuals are to some reference instances, such as the training data instances. We adopt the measure of 'implausibility' from (Guidotti 2022) and formulate the metric of plausibility.

Plausibility of Counterfactual Explanations.

Metric 6. *The* ***plausibility*** *metric* $Q^{PLAUSIBILITY}$ *for measuring soundness of a counterfactual explanation* $E_{\mathbf{x}} = (f_1 : v_1, \ldots, f_{|\mathbf{X}|} : v_{|\mathbf{X}|})$ *given a reference set* $\mathcal{R} \subseteq \prod_{i=1}^{|\mathbf{X}|} D_i$ *assigns to* $E_{\mathbf{x}}$ *its* plausibility score *thus:*

$$Q^{PLAUSIBILITY}(E_{\mathbf{x}}) = 1 - \min_{\mathbf{x}^* \in \mathcal{R}} d(E_{\mathbf{x}}, \mathbf{x}^*). \tag{10}$$

Here, $d : D \times D \longrightarrow [0, 1]$ *is a (normalised) distance function on the domain* $D = \prod_{i=1}^{|\mathbf{X}|} D_i$ *of the (labeless) dataset* X *(see Definition 3.1).*

So plausibility measures how close a counterfactual is to a reference set of data instances, where the reference set captures which situations are possible. The default choice of the reference set can be the dataset $\mathcal{R} = X$ itself (or X_{train} if explanations are generated for X_{test}). In that case, if the counterfactual explanation produced by an

[11] Again, it is called 'feasibility' in (A. Karimi et al. 2023), having in mind actionable/feasible interventions that the receiver of counterfactual explanations is able to perform.

explainer is an actual data instance, i.e. $E_{\mathbf{x}} \in \mathcal{R} = X$, then the counterfactual explanation is entirely plausible with plausibility score $Q^{\text{PLAUSIBILITY}}(E_{\mathbf{x}}) = 1 - d(E_{\mathbf{x}}, E_{\mathbf{x}}) = 1 - 0 = 1$, irrespective of the chosen distance function d. For example, all the counterfactual explanations in $C = \{(1,0), (0,1), (1,1)\}$ from Example 4.5 are entirely plausible given the reference set $\{0,1\}^2$. In general, a higher plausibility score is more desirable. Also, note that the metric can be extended to sets of counterfactual explanations, for instance by taking the average of plausibility scores (cf. (Guidotti 2022)).

To illustrate with DiCE and Baseline explainers, we take the reference set consisting of training instances classified by the model the same as the counterfactual. That is, let $\mathcal{R} = \{\mathbf{x}' \in X_{train} \,:\, \mathcal{M}(\mathbf{x}') = \mathcal{M}(E_{\mathbf{x}})\}$. We use the Euclidean distance $\|\cdot\|_2$ to get similarities $\frac{1}{1+\|E_{\mathbf{x}} - \mathbf{x}^*\|_2}$, which we min-max normalise across all explanations $E_{\mathbf{x}}$ to rescale the similarities to span the full interval $[0,1]$. We use 1 minus the normalised similarity as $d(E_{\mathbf{x}}, \mathbf{x}^*)$, so that the plausibility score of $E_{\mathbf{x}}$ is min-max normalised $\frac{1}{1+\|E_{\mathbf{x}} - \mathbf{x}^*\|_2}$. In Table 6, we report the average plausibility of DiCE and Baseline counterfactual explanations generated for test instances. We note that in any setting of data, model and explainer (as well as any parameters), counterfactual explanations are neither entirely plausible nor implausible, with average plausibility varying greatly across combinations of models and explainers.

Table 6. Average plausibility scores of DiCE and Baseline counterfactual explanations generated with respect to XGB and NN models for instances from test datasets, with respect to the corresponding reference sets consisting of the training instances classified by the model the same as the counterfactual. (λ is the Baseline regularisation parameter.)

Model type	**Dataset ($\lvert X_{test}\rvert$ = #expls)**	DiCE	Baseline				
			$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 10$	$\lambda = 100$
XGB	Iris (35)	–	0.499	0.517	0.503	0.481	0.480
	Spiral (20)	0.591	0.565	0.606	0.542	0.485	0.434
	Telecom (200)	0.457	0.956	0.944	0.941	0.946	0.936
NN	Iris (35)	–	0.516	0.512	0.533	0.520	0.498
	Spiral (20)	0.674	0.571	0.559	0.537	0.528	0.561
	Telecom (200)	0.366	0.958	0.958	0.952	0.936	0.925

We observe that the metric of plausibility as defined above is meant to also capture similar notions such as that of *sparsity* (also called 'minimality' in (Guidotti 2022)) or *data manifold closeness* (Verma et al. 2022). For instance, sparsity quantifies the distance between an instance and its counterfactual, typically measuring how many feature values change and by how much. So in plausibility we could set the reference set to be the singleton $\{\mathbf{x}\}$ of the instance explained, so that the lower the measured sparsity between $\mathbf{x}$ and $E_{\mathbf{x}}$, the higher the plausibility of $E_{\mathbf{x}}$, as desired. Similarly, the counterfactual explanation's adherence to a data manifold defined by some portion of the training dataset can be measured (or even imposed during generation) via distance to the k-nearest data points, or estimation of the data manifold density or other means – see (Verma et al. 2022) for an overview. In the end, such measures should boil down to some form of distance from the counterfactual to the reference set of data points.

Overall, a counterfactual explanation $E_{\mathbf{x}}$ is measured to be sound in as much as it is plausible with respect to a reference set of possibilities, feasible in terms of altering the feature values of $\mathbf{x}$ according to given constraints, and valid as long as the model output for $E_{\mathbf{x}}$ is different from the one for $\mathbf{x}$. We maintain it is highly desirable for counterfactual explanations to be entirely valid, feasible and plausible, with validity $Q^{\text{VALIDITY}}(E_{\mathbf{x}})$, feasibility $Q^{\text{FEASIBILITY}}(E_{\mathbf{x}})$ and plausibility $Q^{\text{PLAUSIBILITY}}(E_{\mathbf{x}})$ scores of 1 – after all, if an explanation says that a change in the input feature values would result into changed model output, then that should be a true statement about possible to obtain circumstances if the explanation is to counterfactually explain the model's behaviour.

Now let us consider what it would mean for counterfactual explanations to be complete.

4.2.2 Completeness of Counterfactual Explanations. We would like to say that counterfactual explanations are complete when each of them explains some input(s) to the model and they collectively explain all the model inputs in a diverse manner. Intuitively, we think of completeness of a set of counterfactual explanations, rather than a single counterfactual. We propose that a collection of counterfactuals is complete in as much as each is an explanation to as many instances and all collectively represent many diverse possible worlds.

Property 4 (Completeness of counterfactual explanations). *A set E of counterfactual explanations is **complete** in as much as it contains diverse counterfactuals to all of the ML model input instances.*

On the one hand, completeness of counterfactual explanations refers to some form of *coverage* of each and all counterfactuals (Keane et al. 2021; Mohammadi et al. 2021). On the other hand, it concerns some form of *diversity* across counterfactuals. Let us see how to measure these aspects.

Coverage can be understood in the sense of the explainer being able to produce counterfactual explanations for all data instances (Mohammadi et al. 2021). While this seems only natural and desirable, not all counterfactual explainers behave this way, e.g. (Mothilal et al. 2020) according to (Mohammadi et al. 2021). Note that this intuitively should apply to sets of, rather than individual counterfactual explanations. Yet, for a single counterfactual explanation, we may still stipulate that it covers instances to which it would be generated as a counterfactual by the explainer, even if was generated for some specific instance. We thus state a new (to the best of our knowledge) coverage metric for counterfactual explanations.

Coverage of Counterfactual Explanations.

Metric 7. *The **coverage** metric $Q^{COVERAGE}$ for measuring completeness of a counterfactual explanation $E_{\mathbf{x}} = \mathsf{EX}(\mathcal{M}(\mathbf{x}) = y)$ or a set $E = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_m\}$ of counterfactual explanations (generated by explainer EX) assigns to $E_{\mathbf{x}}$ or E, respectively, its* coverage value *thus:*

$$Q^{COVERAGE}(E_{\mathbf{x}}) = \frac{|\{\hat{\mathbf{x}} \in X \ : \ \mathsf{EX}(\mathcal{M}(\hat{\mathbf{x}}) = \hat{y}) = E_{\mathbf{x}}\}|}{|X|}; \tag{11}$$

$$Q^{COVERAGE}(E) = \frac{|\{\hat{\mathbf{x}} \in X \ : \ \mathsf{EX}(\mathcal{M}(\hat{\mathbf{x}}) = \hat{y}) = \mathbf{x}' \text{ and } \mathbf{x}' \in E\}|}{|X|}. \tag{12}$$

In other words, counterfactual explanations' coverage is the proportion of instances explained by them. Ideally, a complete set of counterfactual explanations would have full coverage of the dataset X with coverage value 1, so that every input is covered by (at least) one explanation.

Example 4.8. Revisiting Example 4.5, recall the counterfactuals $C = \{\mathbf{x}'_1 = (f_1 : 1, f_2 : 0), \mathbf{x}'_2 = (f_1 : 0, f_2 : 1), \mathbf{x}'_3 = (f_1 : 1, f_2 : 1)\}$. By inspection, they are all counterfactual explanations for input $\mathbf{x} = (0, 0)$ to (the logical disjunction) model $\mathcal{M}$. Indeed, assuming that the explainer EX yields only valid (see Metric 4.2.1) counterfactual explanations, $(0, 0)$ is the only instance that each of them is a counterfactual explanation for. So each of $\mathbf{x}'_i \in E$ has coverage value $\frac{1}{4}$, and collectively $Q^{\text{COVERAGE}}(E) = \frac{1}{4}$ too.

On the other hand, the counterfactual $\mathbf{x}' = (f_1 : 0, f_2 : 0)$ as an explanation for each of the instances $(1, 0), (0, 1), (1, 1)$ has $Q^{\text{COVERAGE}}(\mathbf{x}') = \frac{3}{4}$. So the set $\{\mathbf{x}', \mathbf{x}'_1\}$ of counterfactual explanations has complete coverage of the dataset $\{(0, 0), (1, 0), (0, 1), (1, 1)\}$, with coverage value 1.

In Table 7, we report the coverage scores of DiCE and Baseline counterfactual explanations generated for test instances. In DiCE explainer, we can choose how many counterfactuals to generate for an input instance via parameter total_CFs, and we report cases with 1 and 3 counterfactuals per instance. Since coverage metric requires generating counterfactuals for all the instances from X, this is very computationally costly, especially

for DiCE with more than 1 counterfactual with models trained on Telecom dataset. We note that given 80/20 train/test splits for all datasets, the coverage score of any set of explanations generated from the test set must be at least 0.2, because the explanations trivially cover the instances they explain. Other than that, coverages of explanations vary across data-model-explainer settings, but are mostly relatively low, except for the Baseline counterfactual explanations for models trained on Telecom.

Table 7. Coverage scores of DiCE and Baseline counterfactual explanations generated with respect to XGB and NN models for instances from test datasets. DiCE and DiCE-3 represent sets of counterfactual explanations with 1 and 3 counterfactuals per instance, respectively. The Baseline regularisation parameter λ values 0.1, 1, 5, 10, 100 yield the same plausibility scores.

Model type	**Dataset** ($\lvert X_{test}\rvert$)	DiCE	DiCE-3	Baseline
XGB	Iris (35)	–	–	0.413
	Spiral (20)	0.37	0.53	0.2
	Telecom (200)	0.205	0.205	0.941
NN	Iris (35)	–	–	0.413
	Spiral (20)	0.34	0.48	0.2
	Telecom (200)	0.206	0.206	0.941

Note that instead of looking for data instances $\hat{\mathbf{x}}$ for which the given counterfactual explanation $E_{\mathbf{x}}$ could be generated, we could look for instances to which the given $E_{\mathbf{x}}$ is actually a counterfactual. Namely, we could define the coverage of a counterfactual explanation $E_{\mathbf{x}} = (f_1 : v_1, \ldots, f_{|\mathbf{X}|} : v_{|\mathbf{X}|})$ to be 1 if there is $\hat{\mathbf{x}} = (x_1, \ldots, x_{|\mathbf{X}|}) \in X$ with $x_i \neq v_i$ for some $i \in \{1, \ldots, |\mathbf{X}|\}$ such that $\mathcal{M}(\hat{\mathbf{x}}) \neq \mathcal{M}(E_{\mathbf{x}})$, and 0 otherwise. That is, if $E_{\mathbf{x}}$ differs from some instance $\hat{\mathbf{x}}$ in at least one feature value as well as classification by the model, then it could be deemed a counterfactual to $\hat{\mathbf{x}}$, and thus cover it. However, this would seem to have at least two problems. First, it would be a binary metric of coverage, missing any nuances of counting or proportionality, and thus not very interesting as a measure. Second, this would blend soundness into what is supposed to be a metric for completeness, by mandating what a counterfactual could be, as opposed to leaving that to the explainer. We want to keep the two properties separate and stick to our definition.

Other similar definitions of counterfactual coverage are possible though. For instance, (Keane et al. 2021) define coverage as a user-dependent metric that measures the proportion of counterfactual explanations that actually explain some instance in the dataset, where 'actually explains' is meant to capture some user-dependent notion of counterfactual explanatory power. We do not aim to speculate about what "explaining to a human user" means and stick to a simple computational measure, assuming that any and all counterfactuals generated by an explainer are taken to be explanatory in that sense. We also note that counterfactuals can to begin with be intended to explain groups of instances rather than individual ones (Warren et al. 2023), in which case the proportion of that group could be taken as a counterfactual's coverage. In our metric this would amount to replacing the whole dataset X with the intended reference group, for example only instances with a particular class label.

We next consider diversity of counterfactual explanations. As with data instances, counterfactual instance diversity can be measured in terms of distance among them. Following (Guidotti 2022; Mohammadi et al. 2021; Mothilal et al. 2020), we define the diversity metric via aggregation of pair-wise distances between counterfactuals.

Diversity of Counterfactual Explanations.

Metric 8. *The **diversity** metric $Q^{DIVERSITY}$ for measuring completeness of a set $E = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_\mathbf{m}\}$ consisting of at least 2 counterfactual explanations assigns to E its* diversity value *thus:*

$$Q^{DIVERSITY}(E) = \frac{2 \cdot \sum_{(\mathbf{x}', \mathbf{x}'') \in E \times E} d(\mathbf{x}', \mathbf{x}'')}{|E| \cdot (|E| - 1)}. \tag{13}$$

Here, $d : D \times D \to [0, 1]$ is a distance function on the domain $D = \prod_{i=1}^{|\mathbf{X}|} D_i$ of the (labeless) dataset X (see Definition 3.1).

Note that there are $\frac{|E| \cdot (|E|-1)}{2}$ unordered pairs with distinct elements in a set E. These have possibly non-zero distances (since any distance function d satisfies $d(x, x) = 0$ for any x), and so are the ones that contribute to the sum. By fiat, we can also agree that the empty set $\emptyset$ or any singleton set $\{E_\mathbf{x}\}$ would be assigned diversity value 0.

In other words, the diversity of a set of counterfactual explanations is the averaged pair-wise distance between its elements. Preferably, a more complete set of counterfactual explanations would have higher diversity value, ideally 1 if all its counterfactuals are as distant from each other as possible. For example, using the normalised Hamming distance[12] among counterfactual explanations from Example 4.8, the set $E' \coloneqq \{\mathbf{x}' = (f_1 : 0, f_2 : 0), \mathbf{x}'_1 = (f_1 : 1, f_2 : 0)\}$ would have $Q^{\text{DIVERSITY}}(E') = \frac{1}{2}$ whereas $\{\mathbf{x}' = (f_1 : 0, f_2 : 0), \mathbf{x}'_3 = (f_1 : 1, f_2 : 1)\}$ would have diversity value 1.

To illustrate with DiCE and Baseline explainers, we implement the diversity metric using the (normalised) cosine distance $d(\mathbf{x}', \mathbf{x}'') = (1 - \frac{\mathbf{x}' \cdot \mathbf{x}''}{\|\mathbf{x}'\|_2 \|\mathbf{x}''\|_2})/2$. We report the diversity values of DiCE and Baseline counterfactual explanations generated for test instances in Table 8. We note that the diversity values of either DiCE or Baseline explanations in any setting are at best middling, typically under 0.5, indicating that many of the counterfactuals are quite similar.

Table 8. Diversity values of DiCE and Baseline counterfactual explanations generated with respect to XGB and NN models for instances from test datasets. (λ is the Baseline regularisation parameter.)

Model type	**Dataset** ($\lvert X_{test}\rvert$ = #**expls**)	DiCE	Baseline				
			$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 10$	$\lambda = 100$
XGB	Iris (35)	–	0.354	0.347	0.344	0.373	0.340
	Spiral (20)	0.440	0.475	0.353	0.354	0.255	0.207
	Telecom (200)	0.498	0.053	0.059	0.052	0.031	0.029
NN	Iris (35)	–	0.356	0.349	0.339	0.328	0.358
	Spiral (20)	0.445	0.475	0.404	0.313	0.362	0.525
	Telecom (200)	0.367	0.057	0.058	0.057	0.079	0.080

Overall, we typically measure a set E of counterfactual explanations to be complete in as much as E contains diverse explanations in terms of some distance metric and has high coverage value in terms of explaining as many instances as possible.

We finally consider feature importance-based explanations.

[12] In this case, giving the ratio of the number of feature value flips to obtain one instance from another over the number of feature-value pairs: $d\left(\{f_1 : v_1, \ldots, f_n : v_n\}, \{f_1 : v'_1, \ldots, f_n : v'_n\}\right) = \frac{1}{n} \sum_{1 \leqslant i \leqslant n} [v_i \neq v'_i]$, where $[a \neq b]$ is the Iverson bracket yielding 1 if a does not equal b and 0 otherwise.

4.3 Correctness of Feature Importance-Based Explanations

Intuitively, feature importance-based explanations (see Definition 3.1) are correct if they correctly reflect the impact that features have on the ML model outputs. We will stipulate that 'correct reflection' amounts to some sort of adequacy and representativeness of importance scores. This will amount to soundness and completeness, respectively.

4.3.1 Soundness of Feature Importance-based Explanations. Following (Nauta et al. 2023), feature importance-based explanations reflect how much impact or relevance each feature actually has on the ML model outputs. This could be observed by perturbing the features (e.g. by removing, masking, or changing their values) and observing (proportional) changes to the model outputs; see e.g. (A. Fisher et al. 2019) for an extensive study of variable importance to ML model input/output behaviour. The relevance could otherwise be validated by domain experts with knowledge of causal relationships or statistical patterns that are assumed to have been learned by the model. In any event, changes to features with higher (respectively, lower) importance scores should result into bigger (respectively, smaller) changes to model outputs.

Property 5 (Soundness of feature importance-based explanations). *A (local or global) feature importance-based explanation $E = \{f_1 : s_1, \ldots, f_n : s_n\}$ is **sound** in as much as the feature importance scores $s_1, \ldots, s_n$ adequately reflect the impact or relevance that features $f_1, \ldots, f_n$ have on the underlying ML model outputs.*

Unlike the soundness properties for counterfactual and rule-based explanations above, the one for attribution-based explanations is neither as precise nor as prescriptive. In particular, the notion of 'adequately reflecting' leaves room for interpretation. This is so on purpose because we want to allow for a broader family of ways to quantify feature relevance. In any case, we posit that it is desirable for feature importance-based explanations to be as sound as possible.

We now consider how to measure soundness of feature importance-based explanations. A key is to use input instance perturbations and observe model output shifts. In a nutshell, feature importance scores are expected to be proportional to the shift in output distribution of the underlying ML model applied to inputs with feature values perturbed. To measure this, we next formulate a rather abstract and general metric of explanation *fidelity*, inspired by (Yeh 2019). We will concretise it immediately after.

The basic idea behind fidelity is to quantify the correlation between the feature importance scores applied to a perturbed instance and the shift of the ML model's output when input with the perturbed instance. The starting point is to perturb the given instance $\mathbf{x} = (x_1, \ldots, x_{|\mathbf{X}|}) \in X$ with random perturbations to its feature values x_i, often with respect to some baseline $\mathbf{x}_B \in X$. A perturbation can be simply thought of as changing at least one x_i to some other value from the domain D_i of feature f_i, though desirably in the way that the resulting perturbed instance $\mathbf{x}'$ is meaningful, e.g. at least belongs to the domain $D = \prod_i^{|\mathbf{X}|} D_i$ of X. With a perturbed instance $\mathbf{x}'$, one can measure the shift in the model outputs from $\mathcal{M}(\mathbf{x})$ to $\mathcal{M}(\mathbf{x}')$, and compare that to the shift in inputs from $\mathbf{x}$ to $\mathbf{x}'$, weighted by feature importance scores. For example, (Yeh 2019) specifically measure the expected mean square error between the two shifts. In general, for an explanation to be sound as regards its fidelity, the shift in model outputs should be proportional to the shift in input perturbations weighted by feature importance, in the expectation of varying the perturbations. We thus state the metric of fidelity, inspired by and adopting the measure of explanation infidelity (Yeh 2019, Definition 2.1), as follows.

Fidelity of Feature Importance-based Explanations.

Metric 9. *The **fidelity** metric $Q^{FIDELITY}$ for measuring soundness of a local/global feature importance-based explanation $E = E_{\mathbf{x}}/E_{\mathbf{X}} = \{f_1 : s_1, \ldots, f_n : s_n\}$ assigns to E its* fidelity score *as the expected value of correlation*

between instance perturbations weighted by feature importance and the shift in model outputs thus:

$$Q^{FIDELITY}(E_{\mathbf{x}}) = \mathbb{E}_{\mathbf{v}\sim\mu}\left[corr\left(E_{\mathbf{x}} \otimes \mathbf{v}, sim(\mathcal{M}(\mathbf{x}), \mathcal{M}(\mathbf{x} \oplus \mathbf{v}))\right)\right]; \tag{14}$$

$$Q^{FIDELITY}(E_{\mathbf{X}}) = \frac{1}{|X|}\sum_{\mathbf{x}\in X} \mathbb{E}_{\mathbf{v}\sim\mu}\left[corr\left(E_{\mathbf{X}} \otimes \mathbf{v}, sim(\mathcal{M}(\mathbf{x}), \mathcal{M}(\mathbf{x} \oplus \mathbf{v}))\right)\right]. \tag{15}$$

Here:

$\mathbf{v} \in \prod_{i=1}^{n} D_i \subseteq \prod_{i=1}^{|\mathbf{X}|} D_i = D \supseteq X$ *is a random variable with probability measure* μ *such that* $\mathbf{x} \oplus \mathbf{v} \in X$ *represents a meaningful perturbation of* $\mathbf{x}$*, with* $\oplus = (\oplus_1, \ldots, \oplus_{|\mathbf{X}|})$ *being a feature-wise shift operator on the domain* D *of the (labeless) dataset* X *(see Definition 3.1);*
$sim : Y \times Y \to \mathbb{R}$ *is a similarity measure on the class labels* Y *(or their probability distributions);*
$\otimes : \mathbb{R}^n \times D \to \mathbb{R}^m$*, for* $1 \leqslant m \leqslant |\mathbf{X}|$*, is an operator weighting perturbations* $\mathbf{v} = (v_1, \ldots, v_{|\mathbf{X}|})$ *by feature importance scores* $(s_1, \ldots, s_n)$*;*
$corr$ *is a correlation function.*

We first note that in (Yeh 2019), the metric used is called 'explanation infidelity'. That is because with their particular definition, the higher the expected value, the less the feature importance-based explanation is faithful – i.e. has less impact/relevance – to the underlying ML model. It is thus desirable to have a lesser value of infidelity. With our more abstract definition, whether lower or higher value of fidelity is more desirable, depends on the distance functions involved. For example, if *sim* gives higher values to more similar items, $\otimes$ weighs perturbations proportionally to feature importance scores and *corr* grows as the measured quantities correlate, then we would hope for lower values of Q^{FIDELITY}, because we expect bigger changes to model outcomes with bigger input perturbations. Conversely, if – exclusively – either *sim* measures dissimilarity instead, or *corr* is inversely proportional, then the higher $Q^{\text{FIDELITY}}(E)$ the more faithful E is. For instance, in the concrete instantiation of this metric to be given shortly, *sim* is the absolute difference between probabilities which gives lower values to more similar items, while *corr* is Pearson correlation which grows as the correlations get stronger, so that the fidelity score is directly proportional to the faithfulness of the explanation.

Now let us unpack the definition at a high level. For a local explanation $E_{\mathbf{x}}$ to a given instance $\mathbf{x}$, we consider some meaningful perturbations $\mathbf{v}$ and check how well the perturbation-weighted feature importance scores correlate with the shifts in model $\mathcal{M}$ outputs when the inputs shift from the instance $\mathbf{x}$ explained to its perturbed instances $\mathbf{x} \oplus \mathbf{v}$. With local explanations, we can think of different spaces of perturbations to consider. For instance, we know the predicted class $y = \mathcal{M}(\mathbf{x})$, and we know all the instances in the dataset X classified as y (let $X_y \coloneqq \{\mathbf{x} \in X : \mathcal{M}(\mathbf{x}) = y\}$), and thus know all the perturbations $\mathbf{v}$ to $\mathbf{x}$ that comprise X_y. We can then consider the uniform distribution μ over X_y to compute the fidelity score of $E_{\mathbf{x}}$. Or we could consider some local neighbourhood of $\mathbf{x}$ in X and a distance-based distribution therein. With global explanations, we can instead aggregate the local correlations across the whole dataset. Again, we can think of, say, the closest neighbour to or local neighbourhoods around each $\mathbf{x} \in X$, and average the expected values of correlations over X to compute the fidelity score of $E_{\mathbf{X}}$. And of course the aggregation can be more complex than averaging, for instance, aggregating with some distribution over X. We try not to over-complicate the already complex definition, simply noting that it can be extended in some ways.

As noted, the above definition is abstract and rather complex; in concrete settings it often boils down to a simpler equation. For example, with all the features f_i taking numerical values from intervals $D_i \subseteq \mathbb{R}$, a perturbation $\mathbf{v}$ simply changes (some of the) n values of $\mathbf{x}$ to make a perturbed instance $\mathbf{x}' = \mathbf{x} \oplus \mathbf{v}$; further, assuming $\mathcal{M}$ yields probabilities $\mathcal{M}(\mathbf{x})_y$ for each class label $y \in Y$, *sim* could simply be the absolute difference $|\mathcal{M}(\mathbf{x})_y - \mathcal{M}(\mathbf{x}')_y|$ between the model probabilities for the initially predicted class $y = \mathcal{M}(\mathbf{x})$ before and after the perturbation; the weighting of the perturbed instance by the explanation could amount to the weighted sum $E_{\mathbf{x}} \otimes \mathbf{v} = \sum_{i=1}^{n} s_i|v_i|$ of absolute feature value changes and *corr* could be the Pearson correlation r. Then, sampling $m \in \mathbb{N}$ perturbations

$\mathbf{v_j}$ of $\mathbf{x}$ would approximate the fidelity of $E_\mathbf{x}$ as $r_{\mathbf{v},\mathbf{x}'}\left(\sum_{i=1}^{n} s_i|v_i|, |\mathcal{M}(\mathbf{x})_y - \mathcal{M}(\mathbf{x}'_\mathbf{j})_y|\right)$. Specifically in our empirical setup, we implement the fidelity metric using Pearson correlation on 100 perturbations of a given input instance, with random perturbations for each feature value from a normal (Gaussian) distribution with mean of 0 (i.e. centered around the original value) and standard deviation of 0.1.

In terms of explainers in our empirical setup described in Section 3.4, we use the SHAP explainer thus: for XGB models, we load shap.TreeExplainer; for NN models, we load shap.KernelExplainer also passing the background data of a 100 sampled train instances shap.sample(X_train, 100) with the data parameter. We use the Lime explainer lime.lime_tabular.LimeTabularExplainer passing the whole X_{train} with the training_data parameter. We use the IG explainer captum.attr.IntegratedGradients for NN models and extract from it feature importance scores using the attribute method with n_steps=50 in the numerical approximation of the integral when calculating integrated gradients. Other explainer parameters are default.

We report in Table 9 the average fidelity scores as average positive and negative Pearson correlation values of explanations generated for test instances (recall that IG explanations are only for NN models). In general, the average fidelity is quite low, but never zero, indicating that correlations exist, albeit not strong. They are weakest with respect to models trained on Telecom. SHAP and IG explanations seem to be generally more faithful than Lime, but note that these experimental values are highly dependent on the setting and metric instantiation, and thus at best indicative of trends.

Table 9. Average fidelity scores (as positive and negative Pearson correlations) of SHAP, Lime and IG feature importance-based explanations generated with respect to XGB and NN models for instances from test datasets.

Model type	**Dataset** ($\lvert X_{test}\rvert$ = $\#expls$)	SHAP		LIME		IG	
		Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
XGB	Iris (35)	0.231	-0.213	0.212	-0.212	–	–
	Spiral (20)	0.297	-0.352	0.336	-0.327	–	–
	Telecom (200)	0.238	-0.226	0.158	-0.161	–	–
NN	Iris (35)	0.572	-0.506	0.460	-0.378	0.535	-0.488
	Spiral (20)	0.537	-0.644	0.578	-0.635	0.548	-0.593
	Telecom (200)	0.112	-0.145	0.147	-0.141	0.232	-0.199

In a similarly special case using the *incremental* (feature) *deletion* method (see e.g. (Arya, Bellamy, P. Chen, et al. 2019)), there are n perturbed instances $\mathbf{x}'_\mathbf{i} = (x_1, \ldots, x_{i-1}, x_i^B, x_{i+1}, \ldots, x_{|\mathbf{X}|})$, each with the respective feature f_i value set to some base value $x_i^B \neq x_i$; and $\otimes = (\otimes_1, \ldots, \otimes_n)$ with each $\otimes_i : D \rightarrow \mathbb{R}$ yielding the importance score s_i exactly when $v_i \neq 0$ and 0 otherwise.[13] Then with $(p_1, \ldots, p_n)$ denoting the vector of differences in probabilities $p_i = |\mathcal{M}(\mathbf{x})_y - \mathcal{M}(\mathbf{x}'_\mathbf{i})_y|$, the Pearson correlation coefficient can be computed as $r_{\mathbf{s},\mathbf{p}} = \frac{n\sum_{i=1}^{n} s_i p_i - \sum_{i=1}^{n} s_i \sum_{i=1}^{n} p_i}{\sqrt{n\sum_{i=1}^{n} s_i^2 - \left(\sum_{i=1}^{n} s_i\right)^2}\sqrt{n\sum_{i=1}^{n} p_i^2 - \left(\sum_{i=1}^{n} p_i\right)^2}}$.[14]

Another concrete instantiation of measuring Q^{FIDELITY} uses a *cumulative* (feature) *deletion* method whereby instances are considered with a growing number of the most important feature values perturbed instead. In detail, assume (without the loss of generality) that $f_1, \ldots, f_n$ orders features in decreasing importance, i.e. $|s_1| \geqslant \ldots \geqslant |s_n|$.

[13] E.g. $\otimes_i(\mathbf{v}) = s_i \cdot \sum_{j=1}^{|\mathbf{X}|} |sgn(v_j)|$, where $|sgn(v_j)|$ gives the absolute sign value of $v_j \in D_j \subseteq \mathbb{R}$ (i.e. 1 if $v_j \neq 0$ and 0 otherwise), so that $\otimes_i((0, \ldots, 0, x_i - x_i^B, 0, \ldots, 0)) = s_i \cdot |sgn(x_i - x_i^B)| = s_i$.

[14] This is called 'faithfulness' in (Alvarez-Melis and Jaakkola 2018) and is implemented in, for instance, the AI Explainability 360 Toolkit (Arya, Bellamy, P.-Y. Chen, et al. 2019).

Then consider $k \leqslant n$ perturbed instances $\mathbf{x}'_\mathbf{i} = (x_1^B, \ldots, x_{i-1}^B, x_i^B, x_{i+1}, \ldots, x_{|\mathbf{X}|})$, for $1 \leqslant i \leqslant k$. In other words, in the first perturbed instance $\mathbf{x}'_\mathbf{1}$ perturb the most important feature f_1, in the second one $\mathbf{x}'_\mathbf{2}$ perturb the most and the second most important features f_1, f_2, and so on, for k instances. Then, as above, the Pearson correlation coefficient $r_{\mathbf{s},\mathbf{p}}$ can be assigned as Q^{FIDELITY} value to the k-sized most important part $\{f_1 : s_1, \ldots, f_k : s_k\}$ of the explanation.

Fidelity as formulated above essentially covers the single and incremental (feature) deletion methods for measuring explanation 'correctness' as suggested in (Nauta et al. 2023). They basically amount to perturbing one or more feature values to null or some baseline (such as average) values and checking how the model output changes. Relatedly, the incremental (feature) addition method works the other way round, by starting with all feature values being null/baseline and incrementally perturbing the most important features. We note that these methods can be seen as metrics for the 'output-completeness' property which "evaluates whether the set of important features is sufficient to explain the output of [the underlying] model" (Nauta et al. 2023, p. 20). Similarly, the deletion and preservation checks for evaluating whether "the explanation hold[s] enough information to explain the output of [the underlying] model" (Nauta et al. 2023, p. 21) are output-completeness metrics that measure model output shift after removing (resp. adding) the whole explanation from the (resp. to no) input. We effectively tried to capture the intuition behind all these methods with the above definition of fidelity.

The other four correctness properties summarised in (Nauta et al. 2023) are the model parameter and explanation randomization checks as well as the white box and controlled synthetic data checks. Except for partially the last one, we argue that the other three methods are not relevant to this paper. First, the model parameter randomisation check amounts to perturbing model parameters or weights and expecting the resulting explanations to change. We note that it is not clear how universally applicable or desirable this method is. For instance, changing the training parameters for a tree-based model (such as one produced by XGBoost (T. Chen and Guestrin 2016)) may lead to learning a slightly different ensemble of trees yet with the same feature importance scores. Aside from this issue, the method in any case assumes model retraining, whereas in our work we assume a fixed ML model as given. Similarly, the explanation randomisation check amounts to perturbing in-built explanations and hence the model, and should change the model's output. We do not consider this method for the same reason that we consider only fixed models (as well as explanations generated thereof).

Going further, the white box check method aims to evaluate correctness of explanations by applying the explainer to a white box model whose workings are well understood, thus checking if the explanations are truthful to the known model. However, this method merely checks if the explainer produces explanations truthful to the white box model, and does not really say anything about the correctness of explanations of the (black box) model in question. It is only suggestive that if explanations are incorrect for the white box model, then they may be expected to be incorrect for the model in question; and even then perhaps only in specific situations where the two types of models are of similar kind (such as tree-based). We do not think this is a suitable metric for explanation soundness because it does not evaluate explanations directly with respect to the given ML model.

Finally, the controlled synthetic data check amounts to checking whether the features deemed as important are actually important assuming one has great confidence in *a priori* knowledge of feature importance, such as on synthetic datasets curated with such a property in mind. Similarly to the white box check method, this method only checks if a model trained on specific synthetic data admits sound explanations, but not if the model trained on real data does so too. We do not consider this check as a suitable metric either because it does not evaluate explanations with respect to the ML model trained on the given data.

However, the synthetic data check is somewhat similar to the situation where feature importance can be ascertained by domain experts. Alignment of explanations with domain knowledge is part of the 'coherence' property as stated in (Nauta et al. 2023). Indeed, often an annotated dataset is assumed to contain ground truth about feature importance and is used to measure correlation between feature importance-based explanations and the ground truth (Nauta et al. 2023, p. 6.11). Though domain knowledge can also be attained if, for example, some

causal model or connections underlying the task and data or the ML model construction itself are known or can be assumed, see e.g. (Agarwal et al. 2022; Camburu et al. 2019). Thus, in the spirit of (Property 5), and building upon the feature agreement metric from (Agarwal et al. 2022), we will state two more metrics for measuring soundness of feature importance-based explanations, namely *validity* and *agreement* with respect to the ground truth of feature importance. To that end, we first give an auxiliary definition of the top k most important features in a feature importance-based explanation.

Definition 4.9. For a (local or global) feature importance-based explanation $E = \{f_1 : s_1, \ldots, f_n : s_n\}$ and integer $k \in \mathbb{N}$, the k^{th} **most important feature set** is the set $E^k \coloneqq \{(f, s) \in E \ : \ \text{exist at most } k-1 \text{ pairs } (f', s') \in E \text{ with } |s| < |s'|\}$ of features (with their importance scores) such that there are at most $k-1$ more important features.

Example 4.10. The explanation $E = \{f_1 : 1/2, f_2 : 1/2\}$ from Example 3.2 admits the first most important feature set to equal itself: $E^1 = E$. That is, the two features are equally most important (and exhaust the explanation).

Example 4.11. For an example using the SHAP explainer of XGBoost model classification on Telecom dataset introduced in Section 3.4, consider an explanation given in Table 10 – call it E. There, for $k \in \{1, \ldots, 8\}$, the k^{th}

Table 10. A SHAP feature importance-based explanation generated for the XGB model trained on Telecom dataset, represented as top 8 features (out of 12) in descending order of absolute importance (the other feature importance scores are 0).

Feature	Importance Score
7	-2.0446107
1	-1.7867389
4	1.1722904
12	-0.40025863
3	0.34488305
2	-0.24061479
8	-0.21009174
6	0.06578472

most important feature set E^k consists of the top k feature-importance score pairs in terms of absolute importance scores, e.g. $E^3 = \{(f_7, -2.0446107), (f_1, -1.7867389), (f_4, 1.1722904)\}$.

Now, we want to say that a feature importance-based explanation is valid in proportion to how many of the most important features need to be taken to include all of the pre-selected features. Clearly, at least as many top features need to be considered as there are pre-selected ones. How many more – the metric expresses this as a ratio, as we state next.

Validity of Feature Importance-based Explanations .

Metric 10. *The* ***validity*** *metric $Q^{VALIDITY}$ for measuring soundness of a feature importance-based explanation $E = \{f_1 : s_1, \ldots, f_n : s_n\}$ assigns to E its m^{th}-order* validity score *as the ratio of m selected features $f_1^S, \ldots, f_m^S \in \mathbf{X}$ over the number of k most important features for the minimum k such that E^k contains the m selected features thus:*

$$Q_m^{VALIDITY}(E) = \frac{m}{|E^k|} \text{ for } k = \min\{j \in \mathbb{N} \ : \ (f_i^S, s_i) \in E^j \ \forall i \in \{1, \ldots, m\}\}. \tag{16}$$

Example 4.12 (Example 4.11 continued). Suppose the selected features for the Telecom dataset are f_6 and f_8. Then the 2^{nd}-order validity score of E from Example 4.11 is $\frac{2}{|E^8|} = 1/4$.

The higher the m^{th}-order validity score, the better the explanation, we maintain, because we would like as few as possible most important features in an explanation (i.e. E^k) to capture the truly relevant ones, at least as given by the pre-selection that we take as ground truth. Now, it is interesting to not only find the smallest number k of the most important features that cover the estimated ground truth, but also to consider the ratio of the overlap between the top $k = m$ most important features and the m selected ones. In other words, we would like to measure the *agreement* between the most important features as indicated by the explanation and the features pre-selected as ground truth. We thus state the following metric.

Agreement of Feature Importance-based Explanations.

Metric 11. *The* ***agreement*** *metric $Q^{AGREEMENT}$ for measuring soundness of a feature importance-based explanation $E = \{f_1 : s_1, \ldots, f_n : s_n\}$ assigns to E its m^{th}-order* agreement score *as the ratio of the features in the m^{th} most important feature set E^m over the m selected features $f_1^S, \ldots, f_m^S \in \mathbf{X}$ thus:*

$$Q_m^{AGREEMENT}(E) = \frac{|\{f \in \{f_1^S, \ldots, f_m^S\} \ : \ (f, s) \in E^m \text{ for some } s\}|}{m}. \tag{17}$$

In contrast to the validity metric, the agreement metric fixes the number m of features to consider and checks how many of the m pre-selected features are included in the set of the m most important features of the explanation. Nonetheless, as with the m^{th}-order validity score, we maintain that the higher the m^{th}-order agreement score, the better the explanation, because we would like as many of the truly relevant features to be captured by the most important features of the explanation.

Example 4.13 (Example 4.12 continued). With selected Telecom features f_6 and f_8, the 2^{nd}-order agreement score of E from Example 4.11 is $\frac{|E^2|}{2} = 0$.

Example 4.14. Suppose we deem both features f_1 and f_2 equally important for the behaviour of our model $\mathcal{M}$ capturing the logical OR function of two variables from Example 3.2: let $\{f_1, f_2\}$ be pre-selected features. Then the explanation $E = \{f_1 : 1/2, f_2 : 1/2\}$ from Example 3.2, with $E = E^1$ according to Example 4.10, has 2^{nd} order validity score $Q_1^{\text{VALIDITY}}(E) = 2/2 = 1$ as well as 2^{nd} order agreement score $Q_2^{\text{AGREEMENT}}(E) = 2/2 = 1$.

Note that in the above the features $\{f_1^S, \ldots, f_m^S\}$ deemed *a priori* to be most relevant are assumed to be pre-selected either locally (given the model $\mathcal{M}$ and an instance $\mathbf{x}$) or globally (given $\mathcal{M}$), prior to generating either a local- or global-scope explanation E, respectively. Evidently, the validity and agreement metrics are highly subjective in the sense that the relevant feature pre-selection comes from arguably subjective sources, such as domain experts. While domain expert knowledge can be often highly valuable, it need not be immune to mistakes and biases. Indeed, often the purpose or at least a desirable by-product of ML model training is to uncover unexpected patterns; similarly, a desirable outcome of using feature importance-based explanations is uncovering of unexpectedly important features. So whether validity and agreement so defined are desirable, is a matter of debate. Nonetheless, we state them as metrics that can be functionally assessed given *a priori* expert input rather than being dependent on the user (i.e. receiver) of the explanations. Accordingly, we think that the higher the values of m^{th}-order validity and agreement scores, i.e. the more tightly the most important features capture the ones pre-selected to be relevant by domain experts, the better.

We finally consider completeness of feature importance-based explanations.

4.3.2 Completeness of Feature Importance-based Explanations. We would like to say that feature importance-based explanations are complete in as much as they capture feature relevance to model outputs in a representative way.

Property 6 (Completeness of feature importance-based explanations). *A (local or global) feature importance-based explanation $E = \{f_1 : s_1, \ldots, f_n : s_n\}$ is* ***complete*** *in as much as the feature importance scores*

$s_1, \ldots, s_n$ are representative of the impact or relevance that features $f_1, \ldots, f_n$ have on the underlying ML model outputs.

As with the soundness property (Property 5), the completeness property of feature importance-based explanations is not as prescriptive as those of rule-based explanations. Indeed, the word 'representative' can be interpreted in many ways. We stipulate that whether feature importance scores are representative can be observed by inspecting all possible ways of attributing importance to features and looking for one which identifies when the underlying ML model is on aggregate most sensitive to its inputs. We also posit that it is desirable for feature importance-based explanations to be as complete as possible.

We next look at how to measure completeness of feature importance-based explanations. First, we submit that a measure of feature importance-based explanation completeness should be related to the notion of fidelity that measures the soundness of feature importance-based explanations (see Metric 4.3.1). In a nutshell, while fidelity as a measure of soundness quantifies how the feature importance scores correlate with the shifts in model input-output behaviour, the *representativeness* of the attributed feature importance should quantify the ratio of such correlation against some optimal correlation. For this, we state a metric that is to the best of our knowledge new to the XAI literature.

Representativeness of Feature Importance-based Explanations.

Metric 12. *The **representativeness** metric Q^{REPR} for measuring completeness of a local/global feature importance-based explanation $E = E_{\mathbf{x}}/E_{\mathbf{X}} = \{f_1 : s_1, \ldots, f_n : s_n\}$ assigns to E its* representation value *as the ratio of the expected value of correlation between instance perturbations weighted by feature importance and the shift in model outputs against an optimal such value thus:*

$$Q^{REPR}(E) = \frac{\mathbb{E}_{\mathbf{x}\sim X, \mathbf{v}\sim\mu}\left[corr\left(E \otimes \mathbf{v}, sim(\mathcal{M}(\mathbf{x}), \mathcal{M}(\mathbf{x} \oplus \mathbf{v}))\right)\right]}{\text{opt}_{(s'_1,\ldots,s'_n)\in\mathbb{R}^n}\mathbb{E}_{\mathbf{x}\sim X, \mathbf{v}\sim\mu}\left[corr\left(\{f_1 : s'_1, \ldots, f_n : s'_n\} \otimes \mathbf{v}, sim(\mathcal{M}(\mathbf{x}), \mathcal{M}(\mathbf{x} \oplus \mathbf{v}))\right)\right]}. \tag{18}$$

Here: $\mathbf{x}$ is either the instance to which a local explanation $E_{\mathbf{x}}$ is generated, or $\mathbf{x}$ is drawn from X in case a global explanation $E_{\mathbf{X}}$ is generated; $\mathbf{v}$ is drawn from μ so that $\mathbf{x} \oplus \mathbf{v} \in X$ is a perturbation of $\mathbf{x}$, with a feature-wise shift operator $\oplus$, similarity measure sim, perturbation-weighting operator $\otimes$ and correlation function corr as given for the fidelity metric of feature importance-based explanations (Metric 4.3.1); and opt *being either* max *or* min *operator depending on whether, respectively, higher or lower values of correlation mean higher or lower fidelity of the vector $(s'_1, \ldots, s'_n)$ of importance scores. (We also assume the denominator is never zero, e.g. by adding an arbitrarily small ε if needed.)*

The representativeness metric basically measures the ratio of explanation's fidelity against the maximum fidelity over possible feature importance vectors. Similarly as with the fidelity metric, the expected values of correlations depend on which instances and perturbations are considered. With a local-scope explanation $E_{\mathbf{x}}$, we have a fixed instance $\mathbf{x}$ and can vary its neighbourhood and distribution μ therein. With a global-scope explanation $E_{\mathbf{X}}$, we may draw instances $\mathbf{x}$ from the whole dataset X. Intuitively, representation value tells how well the feature importance scores given by the explanation represent feature importance compared to all possible feature importance attributions. Depending on whether higher or lower values of correlation mean higher or lower fidelity, higher or lower $Q^{\text{REPR}}(E)$ is desirable for a feature importance-based explanation E.

Admittedly, the representativeness metric is idealised in the sense that computing it can be computationally prohibitive due to the optimality condition in the denominator of Equation 18. In practice the optimal expected correlation value could be approximated by sampling, similarly to how e.g. Shapley values are approximated (Aas et al. 2021). This can still be prohibitively costly. Instead, we used the fact that we have more than 1 explainer

for any explainer and can thus used the highest *fidelity* value of an explanation as a witness for optimisation. Comparing the fidelity value of an explanation from a given explainer with the maximal fidelity value among explanations from all explainers does not give an ideal representation value, but it is at least a lower bound.

We use the Pearson correlation implementation described above to find, per test instance, the best positive and negative Pearson correlations of explanations, for each class label. We then compute the representation value of an explanation as the ratio between its fidelity score and the highest fidelity score among explanations for the same instance. We report in Table 11 the average representation values of explanations generated for test instances, for each class label. We note that no average representation value is 1. Thus, by comparing feature importance-based explanation fidelity across explainers as in our implementation, we find that the correlations of explanations can still be optimised. And in some case quite a lot: for example, for SHAP explanations of the NN model trained on Telecom.

Table 11. Representation values of SHAP, Lime and IG feature importance-based explanations generated with respect to XGB and NN models for instances from test datasets, for each class label.

Model type	**Dataset (class)**	SHAP	LIME	IG
XGB	Iris (0)	0.815	0.864	–
	Iris (1)	0.886	0.824	–
	Iris (2)	0.876	0.921	–
	Spiral (0)	0.889	0.844	–
	Spiral (1)	0.837	0.921	–
	Telecom (0)	0.918	0.784	–
	Telecom (1)	0.914	0.779	–
NN	Iris (0)	0.893	0.704	0.908
	Iris (1)	0.899	0.430	0.962
	Iris (2)	0.951	0.798	0.970
	Spiral (0)	0.895	1.000	0.872
	Spiral (1)	0.890	0.988	0.858
	Telecom (0)	0.747	0.804	0.830
	Telecom (1)	0.703	0.787	0.865

5 Conclusions

We maintain that it is a fundamental requirement of explanations to be correct, i.e. sound and complete, as much as possible when measured by the relevant metrics. We have proposed three formulations of each of the soundness and completeness properties for three forms of model input/output behaviour explanations, namely rule-based, counterfactuals and feature-importance based. We have presented formal, generic metrics for quantitatively assessing each of the property formulations (Properties 1 to 6). For measuring soundness: Fidelity 4.1.1 for rules; Validity 4.2.1, Feasibility 4.2.1 and Plausibility 4.2.1 for counterfactuals; Fidelity 4.3.1, Validity 4.3.1 and Agreement 4.3.1 for feature importance. For measuring completeness: Coverage 4.1.2 and Representativeness 4.1.2 for rules; Coverage 4.2.2 and Diversity 4.2.2 for counterfactuals; Representativeness 4.3.2 for feature importance. We discussed how each of those metrics is either adopted from or inspired by the relevant measures in the literature, or if our proposed metric is new, and how the formalised metrics aim to encompass the existing measures of explanation assessment, where available.

It seems that the notion of soundness, which addresses how truthful explanations are with respect to the ML model explained, has been well-studied in the literature. This is perhaps it is more natural to require explanations to give "nothing but the truth". Consequently, we have encountered more metrics for measuring this aspect of either of the forms of explanations. In terms of completeness, which addresses how well explanations generalise to explain any of the ML model's behaviours, it seems to have been studied less. This is perhaps not surprising either, assuming that it is less straightforward to define what it means to give "the whole truth". We have thus proposed three novel metrics (Representativeness 4.1.2, Coverage 4.2.2 and Representativeness 4.3.2 – one for each form of explanation) for quantifying completeness of explanations.

We note that our notion of soundness for explanations of ML model input/output behaviour is very similar to the 'output-completeness' property that pertains to "how well the explanation method agrees with the predictions to the original predictive model", proposed in (Nauta et al. 2023, pp. 11, 21, 22). Indeed, the metrics for measuring output-completeness delineated in (Nauta et al. 2023) do overlap with those that we assert for measuring soundness. Notably, we departed from those metrics in many ways, as discussed in Section 4. Further, as with all the explanation quality properties addressed in (Nauta et al. 2023), output-completeness is vaguely worded, whereas we tried to formalise soundness more precisely, at least for the three forms of explanations that we consider. Relatedly, the companion 'reasoning-completeness' property in (Nauta et al. 2023) pertains to explanations describing the internal dynamics of the underlying ML model, which ranges from completely opening up the model and its parameters to training a shadow model that input/output matches the underlying model. That property is not really evaluated quantitatively, but rather qualitatively, whereas our completeness properties are accompanied with quantitative metrics. Importantly, we have submitted formal definitions of each and every metric for measuring all the formulations of our soundness and completeness properties for assessing correctness of explanations, in contrast to the less formal, more summary-type-of overviews of metrics for assessing explanation quality discussed in related work (Section 2).

5.1 Limitations and Future Work

We consider our work here as a step towards formulating desirable properties of explanations and formalising metrics for quantitatively assessing satisfaction of such properties. Our work is limited in several aspects. First and foremost, we focused on the *correctness* of explanations. The way we interpreted it via soundness and completeness turned out rather broad, as witnessed by subsumption of various instantiations of metrics for quantifying correctness. However, the landscape of properties and metrics is arguably much broader, as discussed in Section 2. One could further identify and study the following aspects of explanations: *robustness*, which intuitively pertains pertaining to how stable explanations are; *complexity*, which intuitively pertains to both computational and cognitive costs of generating and consuming explanations; *contextual relevance*, which intuitively pertains to adhering to the context of the explaining process, primarily concerning intelligibility and usability of the explanations to the user (i.e. receiver of the explanations). We believe it is essential to complement the functionally-grounded evaluation of explanation correctness and other aspects with how human users consume explanations. We agree with Miller (Miller 2019) in that most of the research and practice around developing explanation methods is primarily based on the researchers' perspectives on characterising a good explanation and that this could lead to failures of XAI as a field. However, we believe more than anything else that we first need to properly devise computational measures of explanation goodness, as we have attempted here with explanation correctness, and only then turn to human evaluation aspects. It would thus be great to conduct similar formalisation studies as this one wits respect to the above mentioned aspects of explanations in the future.

We are also limited here in not carefully showing how various metrics proposed in XAI literature are formally instances of some of our generalised metrics. This was not our intention in any case, because properly capturing

a variety of specific but conceptually similar measures under any one metric would risk making the metrics too open ended and not so readily applicable. We are perhaps somewhat guilty of this, particularly with the fidelity metric for measuring soundness of feature importance-based explanations, but even there we adopted the general metric from an earlier work and explained how different common instantiations come about. In general, we argued informally that some of our metric formalisations either evidently stem from or cover notable metrics from the literature. It would be an interesting exercise to try to instantiate as many as possible known measures as our metrics, but we leave this for future work.

Perhaps the biggest limitation is the lack of an extensive experimentation and evaluation of different explainers and their explanations, with respect to a variety of ML models and tabular datasets. It would be a next big step to carry out such a study, aiming to empirically establish which explainer produces more correct explanations and in which settings. Some work has attempted that, at least in restricted settings, notably (Agarwal et al. 2022). It is however important to note that such engineering efforts are rather difficult, because in our experience it is far from trivial to integrate a variety of explainers to work with a fixed model and dataset, to make sure they produce explanations in the same form and to implement the metrics thereof. Even our current experimental setup with six off-the-shelf explainers and six ML models proved to be challenging to implement, due to limited library and method cross-compatibility, data processing nuances and the sheer computation time. Our focus in this work was on the conceptual and theoretical development and we are leaving further extensive engineering and experimentation for future work.

Relatedly, it would be interesting to conduct experiments with human users whereby they are able to see how good one or another explanation (or explainer) is in terms of correctness. We have actually carried out, but are not reporting, some such preliminary experiments. To that end, we have implemented a visually interactive tool that shows to the user the metric scores of different explanations generated by various explainers for a given model. We then equipped the system with automated aggregation of metric scores based on the user preferences over properties or metrics of explanations. In particular, the user is able express ordinal preferences over properties or metrics of interest, whence they induce a weighted sum of the metric scores thus yielding an overall score of the 'goodness' of an explanation (in terms of correctness and other aspects). The preliminary experiments indicated that such personalisation of the explaining process seems to make it fairer and useful to the user. We are planning to continue our investigations along these lines as long as we can formalise the computational aspects.

Finally, some other limitations pertain to our setting in this paper. These include: other modalities of data, such as images, text or time series, not limited to tabular data; other forms of explanations, such as graph- or model-based; other tasks than classification, such as regression, unsupervised clustering, self-supervised text generation. We made a remark in Section 3 as to why we believe the current framework could be extended to other settings, bu we leave any such investigations for future work. More interesting for us would be to go beyond input/output explanations of model behaviour, towards *explainable training* – namely, what concepts and how does a model learn while being trained?[15] For instance, extracting explanations of input/output behaviour after each epoch of training (say, a boosted tree or a neural network) in the hope of capturing the concepts learned. Perhaps property- and metric-based analysis could be useful in such setting too, in that one would stipulate how the model should learn and measure whether it does so. We leave these speculative ideas for the future.

Acknowledgements

Vandita Singh and **Kristijonas Čyras** have contributed equally. **Kristijonas Čyras:** This author has been affiliated with Ericsson, Inc. when researching and writing the article presented herein, but at the time of writing this paper is no longer affiliated with Ericsson. The views and opinions of the author expressed herein are personal

[15]This can be seen as (part of) *developmental interpretability*, which studies how structure forms in ML models, see e.g. https://www.lesswrong.com/s/SfFQE8DXbgkjk62JK.

and do not necessarily reflect those of the European Commission or other EU institutions. **Muhammad Zain Akram**: This author contributed to this article as an MSc thesis intern at Ericsson, but at the time of writing is no longer affiliated with this institution.

Author Contributions. **Vandita Singh:** Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation, Visualization, Writing - Review & Editing **Kristijonas Čyras:** Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation, Writing - Review & Editing, Visualization. **Muhammad Zain Akram**: Validation, Investigation. **Rafia Inam:** Funding Acquisition, Resources, Writing - Review.

References

S. A. and S. R.. 2023. "A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends." *Decision Analytics Journal*, 7, 100230. doi:https://doi.org/10.1016/j.dajour.2023.100230.

K. Aas, M. Jullum, and A. Løland. 2021. "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values." *Artif. Intell.*, 298, 103502. doi:10.1016/J.ARTINT.2021.103502.

A. Adadi and M. Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access*, 6, 52138–52160. doi:10.1109/ACCESS.2018.2870052.

C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, and H. Lakkaraju. 2022. "OpenXAI: Towards a Transparent Evaluation of Model Explanations." In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. http://papers.nips.cc/paper%5C_files/paper/2022/hash/65398a0eba88c9b4a1c38ae405b125ef-Abstract-Datasets%5C_and%5C_Benchmarks.html.

E. Albini, J. Long, D. Dervovic, and D. Magazzeni. June 2022. "Counterfactual Shapley Additive Explanations." In: *2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). ACM, (June 2022). doi:10.1145/3531146.3533168.

D. Alvarez-Melis and T. S. Jaakkola. 2018. "Towards Robust Interpretability with Self-Explaining Neural Networks." *CoRR*, abs/1806.07538. http://arxiv.org/abs/1806.07538 arXiv: 1806.07538.

S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling. 2019. "Explainable Agents and Robots: Results from a Systematic Literature Review." In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*. Ed. by E. Elkind, M. Veloso, N. Agmon, and M. E. Taylor. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088. http://dl.acm.org/citation.cfm?id=3331806.

A. B. Arrieta et al.. 2020. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Inf. Fusion*, 58, 82–115. doi:10.1016/J.INFFUS.2019.12.012.

V. Arya, R. K. E. Bellamy, P.-Y. Chen, et al.. Sept. 2019. *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*. (Sept. 2019). https://arxiv.org/abs/1909.03012.

V. Arya, R. K. E. Bellamy, P. Chen, et al.. 2019. "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques." *CoRR*, abs/1909.03012. http://arxiv.org/abs/1909.03012 arXiv: 1909.03012.

U. Bhatt, A. Weller, and J. M. F. Moura. 2021. "Evaluating and Aggregating Feature-Based Model Explanations." In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (IJCAI'20) Article 417. Yokohama, Yokohama, Japan, 7 pages. ISBN: 9780999241165.

F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo. 2023. "Benchmarking and survey of explanation methods for black box models." *Data Min. Knowl. Discov.*, 37, 5, 1719–1778. doi:10.1007/S10618-023-00933-9.

N. Burkart and M. F. Huber. May 2021. "A Survey on the Explainability of Supervised Machine Learning." *J. Artif. Int. Res.*, 70, (May 2021), 245–317. doi:10.1613/jair.1.12228.

O.-M. Camburu, E. Giunchiglia, J. Foerster, T. Lukasiewicz, and P. Blunsom. 2019. "Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods." In: *NeurIPS 2019 Workshop on Safety and Robustness in Decision Making*. Vancouver. http://arxiv.org/abs/1910.02065 arXiv: 1910.02065.

D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. 2019. "Machine Learning Interpretability: A Survey on Methods and Metrics." *Electronics*, 8, 8. doi:10.3390/electronics8080832.

T. Chakraborti, S. Sreedharan, and S. Kambhampati. 2020. "The Emerging Landscape of Explainable Automated Planning & Decision Making." In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. Ed. by C. Bessiere. ijcai.org, 4803–4811. doi:10.24963/IJCAI.2020/669.

Ed. by B. Braunschweig and M. Ghallab. "Trustworthy AI." *Reflections on Artificial Intelligence for Humanity*. Springer International Publishing, Cham, 13–39. ISBN: 978-3-030-69128-8. doi:10.1007/978-3-030-69128-8_2.

T. Chen and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. Ed. by B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi. ACM, 785–794. doi:10.1145/2939672.2939785.

G. Ciatto, F. Sabbatini, A. Agiollo, M. Magnini, and A. Omicini. 2024. "Symbolic Knowledge Extraction and Injection with Sub-symbolic Predictors: A Systematic Literature Review." *ACM Comput. Surv.*, 56, 6, 161:1–161:35. doi:10.1145/3645103.

G. Ciatto, M. I. Schumacher, A. Omicini, and D. Calvaresi. 2020. "Agent-Based Explanations in AI: Towards an Abstract Framework." In: *Explainable, Transparent Autonomous Agents and Multi-Agent Systems - Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9-13, 2020, Revised Selected Papers* (Lecture Notes in Computer Science). Ed. by D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling. Vol. 12175. Springer, 3–20. doi:10.1007/978-3-030-51924-7_1.

K. Čyras, R. Badrinath, S. K. Mohalik, A. Mujumdar, A. Nikou, A. Previti, V. Sundararajan, and A. V. Feljan. 2021. "Machine Reasoning Explainability." In: *AAMAS: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom. https://underline.io/lecture/18444-tutorial-machine-reasoning-explainability*. AAMAS. https://underline.io/lecture/18444-tutorial-machine-reasoning-explainability.

K. Čyras, D. Letsios, R. Misener, and F. Toni. 2019. "Argumentation for Explainable Scheduling." In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2752–2759. doi:10.1609/AAAI.V33I01.33012752.

K. Čyras, A. Rago, E. Albini, P. Baroni, and F. Toni. 2021. "Argumentative XAI: A Survey." In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*. Ed. by Z. Zhou. ijcai.org, 4392–4399. doi:10.24963/IJCAI.2021/600.

W. Ding, M. Abdel-Basset, H. Hawash, and A. M. Ali. 2022. "Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey." *Information Sciences*, 615, 238–292. doi:https://doi.org/10.1016/j.ins.2022.10.013.

F. Doshi-Velez and B. Kim. 2017. "A Roadmap for a Rigorous Science of Interpretability." *CoRR*, abs/1702.08608. http://arxiv.org/abs/1702.08608 arXiv: 1702.08608.

European Commission and Directorate-General for Communications Networks, Content and Technology. 2019. *Ethics guidelines for trustworthy AI*. Publications Office. doi:doi/10.2759/346720.

J. Fandinno and C. Schulz. 2019. "Answering the "why" in answer set programming - A survey of explanation approaches." *Theory Pract. Log. Program.*, 19, 2, 114–203. doi:10.1017/S1471068418000534.

A. Fisher, C. Rudin, and F. Dominici. 2019. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *J. Mach. Learn. Res.*, 20, 177:1–177:81. http://jmlr.org/papers/v20/18-760.html.

R. A. Fisher. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of eugenics*, 7, 2, 179–188.

S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. 2021. "The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making." *Commun. ACM*, 64, 4, 136–143. doi:10.1145/3433949.

Future of Life Institute. 2023. *Foresight in AI Regulation Open Letter*. (2023). https://futureoflife.org/open-letter/foresight-in-ai-regulation-open-letter/.

M. L. Ginsberg. 1986. "Counterfactuals." *Artif. Intell.*, 30, 1, 35–79. doi:10.1016/0004-3702(86)90067-6.

M. Graziani et al.. Apr. 2023. "A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences." English. *Artificial Intelligence Review*, 56, (Apr. 2023), 3473–3504. doi:10.1007/s10462-022-10256-8.

R. Guidotti. 2022. "Counterfactual explanations and how to find them: literature review and benchmarking." *Data Mining and Knowledge Discovery*. doi:https://doi.org/10.1007/s10618-022-00831-6.

A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M. M. Höhne. 2023. "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond." *Journal of Machine Learning Research*, 24, 34, 1–11. http://jmlr.org/papers/v24/22-0142.html.

R. Ibrahim and M. O. Shafiq. Feb. 2023. "Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions." *ACM Comput. Surv.*, 55, 10, Article 206, (Feb. 2023), 37 pages. doi:10.1145/3563691.

A. Ignatiev. 2020. "Towards Trustable Explainable AI." In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. Ed. by C. Bessiere, 5154–5158. doi:10.24963/IJCAI.2020/726.

A. Ignatiev, M. C. Cooper, M. Siala, E. Hebrard, and J. Marques-Silva. 2020. "Towards Formal Fairness in Machine Learning." In: *Principles and Practice of Constraint Programming - 26th International Conference, CP 2020, Louvain-la-Neuve, Belgium, September 7-11, 2020, Proceedings* (Lecture Notes in Computer Science). Ed. by H. Simonis. Vol. 12333. Springer, 846–867. doi:10.1007/978-3-030-58475-7_49.

A. Ignatiev, Y. Izza, P. J. Stuckey, and J. Marques-Silva. 2022. "Using MaxSAT for Efficient Explanations of Tree Ensembles." In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 3776–3785. doi:10.1609/AAAI.V36I4.20292.

A. Karimi, G. Barthe, B. Schölkopf, and I. Valera. 2023. "A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations." *ACM Comput. Surv.*, 55, 5, 95:1–95:29. doi:10.1145/3527848.

A.-H. Karimi, B. Schölkopf, and I. Valera. 2020. *Algorithmic Recourse: from Counterfactual Explanations to Interventions*. (2020). https://doi.org/10.48550/arXiv.2002.06278.

D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi. Jan. 2022. "Trustworthy Artificial Intelligence: A Review." *ACM Comput. Surv.*, 55, 2, Article 39, (Jan. 2022), 38 pages. doi:10.1145/3491209.

M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth. 2021. "If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques." In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*. Ed. by Z. Zhou. ijcai.org, 4466–4474. doi:10.24963/IJCAI.2021/609.

D. P. Kingma and J. Ba. 2015. "Adam: A Method for Stochastic Optimization." In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. http://arxiv.org/abs/1412.6980.

J. Klaise, A. V. Looveren, G. Vacanti, and A. Coca. 2021. "Alibi Explain: Algorithms for Explaining Machine Learning Models." *Journal of Machine Learning Research*, 22, 181, 1–7. http://jmlr.org/papers/v22/21-0017.html.

N. Kokhlikyan et al.. 2020. "Captum: A unified and generic model interpretability library for PyTorch." *CoRR*, abs/2009.07896. https://arxiv.org/abs/2009.07896 arXiv: 2009.07896.

S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju. 2022. *The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective*. (2022). arXiv: 2202.01602 [cs.LG].

T. Kulesza, S. Stumpf, M. M. Burnett, S. Yang, I. Kwan, and W. Wong. 2013. "Too much, too little, or just right? Ways explanations impact end users' mental models." In: *2013 IEEE Symposium on Visual Languages and Human Centric Computing, San Jose, CA, USA, September 15-19, 2013*. Ed. by C. Kelleher, M. M. Burnett, and S. Sauer. IEEE Computer Society, 3–10. doi:10.1109/VLHCC.2013.6645235.

H. Lakkaraju, R. Caruana, E. Kamar, and J. Leskovec. 2019. "Faithful and customizable explanations of black box models." *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 131–138. ISBN: 9781450363242. doi:10.1145/3306618.3314229.

T. Laugel, M. J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. 2019. "The dangers of post-hoc interpretability: Unjustified counterfactual explanations." *IJCAI International Joint Conference on Artificial Intelligence*, 2019-Augus, 2801–2807. ISBN: 9780999241141. arXiv: 1907.09294. doi:10.24963/ijcai.2019/388.

P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. 2021. "Explainable AI: A Review of Machine Learning Interpretability Methods." English. *Entropy (Basel, Switzerland)*, 23, 1, 18.

S. M. Lundberg and S. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

T. Miller. 2019. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence*, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

T. Miller, P. Howe, and L. Sonenberg. 2017. *Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*. (2017). arXiv: 1712.00547 [cs.AI].

S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. 2021. "Algorithmic Fairness: Choices, Assumptions, and Definitions." *Annual Review of Statistics and Its Application*, 8, Volume 8, 2021, 141–163. doi:https://doi.org/10.1146/annurev-statistics-042720-125902.

K. Mohammadi, A. Karimi, G. Barthe, and I. Valera. 2021. "Scaling Guarantees for Nearest Counterfactual Explanations." In: *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*. Ed. by M. Fourcade, B. Kuipers, S. Lazar, and D. K. Mulligan. ACM, 177–187. doi:10.1145/3461702.3462514.

R. K. Mothilal, A. Sharma, and C. Tan. Jan. 2020. "Explaining machine learning classifiers through diverse counterfactual explanations." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, (Jan. 2020). https://doi.org/10.48550/arXiv.1905.07697.

M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert. 2023. "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI." *ACM Comput. Surv.*, 55, 13s, 295:1–295:42. doi:10.1145/3583558.

I. Nunes and D. Jannach. 2017. "A systematic review and taxonomy of explanations in decision support and recommender systems." *User Model. User Adapt. Interact.*, 27, 3-5, 393–444. doi:10.1007/S11257-017-9195-0.

F. Orilia and M. Paolini Paoletti. 2022. "Properties." In: *The Stanford Encyclopedia of Philosophy*. (Spring 2022 ed.). Ed. by E. N. Zalta. Metaphysics Research Lab, Stanford University.

A. Paszke et al.. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, 8024–8035. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

F. Pedregosa et al.. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825–2830.

P. J. Phillips, C. Hahn, P. Fontana, A. Yates, K. K. Greene, D. Broniatowski, and M. A. Przybocki. Sept. 2021. *Four Principles of Explainable Artificial Intelligence*. en. (Sept. 2021). doi:https://doi.org/10.6028/NIST.IR.8312.

G. Ras, N. Xie, M. van Gerven, and D. Doran. 2022. "Explainable Deep Learning: A Field Guide for the Uninitiated." *J. Artif. Intell. Res.*, 73, 329–396. doi:10.1613/JAIR.1.13200.

S. Raschka. Apr. 2018. "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack." *The Journal of Open Source Software*, 3, 24, (Apr. 2018). https://joss.theoj.org/papers/10.21105/joss.00638.

M. T. Ribeiro, S. Singh, and C. Guestrin. 2016a. ""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* (KDD '16). Association for Computing Machinery, San Francisco, California, USA, 1135–1144. ISBN: 9781450342322. doi:10.1145/2939672.2939778.

M. T. Ribeiro, S. Singh, and C. Guestrin. 2016b. *Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance*. (2016). arXiv: 1611.05817 (stat.ML).

M. T. Ribeiro, S. Singh, and C. Guestrin. 2018. "Anchors: High-Precision Model-Agnostic Explanations." In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by S. A. McIlraith and K. Q. Weinberger. AAAI Press, 1527–1535. doi:10.1609/AAAI.V32I1.11491.

M. Robnik-Šikonja and M. Bohanec. 2018. "Perturbation-Based Explanations of Prediction Models." In: *Human and Machine Learning - Visible, Explainable, Trustworthy and Transparent*. Human-Computer Interaction Series. Ed. by J. Zhou and F. Chen. Springer, 159–175. doi:10.1007/978-3-319-90403-0_9.

A. Rosenfeld. 2021. "Better Metrics for Evaluating Explainable Artificial Intelligence." In: *Adaptive Agents and Multi-Agent Systems*. https://api.semanticscholar.org/CorpusID:233453690.

W. Samek, G. Montavon, A. Binder, S. Lapuschkin, and K.-R. Müller. 2016. *Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation*. (2016). arXiv: 1611.08191 (stat.ML).

G. Schwalbe and B. Finzel. Jan. 2023. "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts." *Data Mining and Knowledge Discovery*, (Jan. 2023). doi:10.1007/s10618-022-00867-8.

R. Schwartz, A. Vassilev, K. K. Greene, L. Perine, A. Burt, and P. Hall. Mar. 2022. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. en. (Mar. 2022). doi:https://doi.org/10.6028/NIST.SP.1270.

T. Shaikhina, U. Bhatt, R. Zhang, K. Georgatzis, A. Xiang, and A. Weller. 2021. "Effects of Uncertainty on the Quality of Feature Importance Estimates." In: *AAAI Workshop on Explainable Agency in Artificial Intelligence*. https://umangsbhatt.github.io/reports/AAAI_XAI_QB.pdf.

V. Singh. 2021. "Explainable AI Metrics and Properties for Evaluation and Analysis of Counterfactual Explanations." Master's thesis. Uppsala University, Department of Information Technology, Uppsala University, Department of Information Technology, 74.

V. Singh, K. Čyras, and R. Inam. 2022. "Explainability Metrics and Properties for Counterfactual Explanation Methods." In: *Explainable and Transparent AI and Multi-Agent Systems*. Ed. by D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling. Springer International Publishing, Cham, 155–172. ISBN: 978-3-031-15565-9.

K. Sokol and P. Flach. 2020. "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* '20). Association for Computing Machinery, Barcelona, Spain, 56–67. ISBN: 9781450369367. doi:10.1145/3351095.3372870.

M. Sundararajan, A. Taly, and Q. Yan. Aug. 2017. "Axiomatic Attribution for Deep Networks." In: *Proceedings of the 34th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Ed. by D. Precup and Y. W. Teh. Vol. 70. PMLR, (Aug. 2017), 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html.

W. R. Swartout, C. Paris, and J. D. Moore. 1991. "Explanations in Knowledge Systems: Design for Explainable Expert Systems." *IEEE Expert*, 6, 3, 58–64. doi:10.1109/64.87686.

A. Terra, R. Inam, S. Baskaran, P. Batista, I. Burdick, and E. Fersman. 2020. "Explainability Methods for Identifying Root-Cause of SLA Violation Prediction in 5G Network." In: *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 1–7. doi:10.1109/GLOBECOM42002.2020.9322496.

S. Upadhyay, S. Joshi, and H. Lakkaraju. 2021. *Towards Robust and Reliable Algorithmic Recourse*. (2021). arXiv: 2102.13620 (cs.LG).

S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah. 2022. *Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review*. (2022). https://doi.org/10.48550/arXiv.2010.10596.

S. Wachter, B. Mittelstadt, and C. Russell. 2018. *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. (2018). https://doi.org/10.48550/arXiv.1711.00399.

G. Warren, M. T. Keane, C. Guéret, and E. Delaney. 2023. "Explaining Groups of Instances Counterfactually for XAI: A Use Case, Algorithm and User Study for Group-Counterfactuals." *CoRR*, abs/2303.09297. arXiv: 2303.09297. doi:10.48550/ARXIV.2303.09297.

C.-k. Yeh. 2019. "On the (In)fidelity and Sensitivity of Explanations." *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, NeurIPS.

J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger. 2021. "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics." *Electronics*, 10, 5. doi:10.3390/electronics10050593.

Received 24 March 2025; accepted 25 June 2025