XAI Handbook: Towards a Unified Framework for Explainable AI

Sebastian Palacio* †‡ Adriano Lucieri* †‡ Mohsin Munir†‡ Sheraz Ahmed†
Jörn Hees† Andreas Dengel†‡
†German Research Center for Artificial Intelligence (DFKI)
‡TU Kaiserslautern
first.last@dfki.de

Abstract

The field of explainable AI (XAI) has quickly become a thriving and prolific community. However, a silent, recurrent and acknowledged issue in this area is the lack of consensus regarding its terminology. In particular, each new contribution seems to rely on its own (and often intuitive) version of terms like "explanation" and "interpretation". Such disarray encumbers the consolidation of advances in the field towards the fulfillment of scientific and regulatory demands e.g., when comparing methods or establishing their compliance w.r.t. biases and fairness constraints. We propose a theoretical framework that not only provides concrete definitions for these terms, but it also outlines all steps necessary to produce explanations and interpretations. The framework also allows for existing contributions to be re-contextualized such that their scope can be measured, thus making them comparable to other methods. We show that this framework is compliant with desiderata on explanations, on interpretability and on evaluation metrics. We present a use-case showing how the framework can be used to compare LIME, SHAP and MDNet, establishing their advantages and shortcomings. Finally, we discuss relevant trends in XAI as well as recommendations for future work, all from the standpoint of our framework.

1. Introduction

The growing demand for explainable methods in artificial intelligence (a.k.a. *eXplainable AI* or XAI) has recently caused a large influx of research on the subject [3].However, proposed solutions have spanned into multiple niches simply because they are based on different definitions for terms like "explanation" or "interpretation". An increasing number of contributions rely on their own, and often intuitive notions of such terms –a phenomenon dubbed "the inmates running the asylum"–which eventually leads to failure to provide satisfactory explanations [26].

This self-reliance in the goal's definition has caused confusion in the machine learning community, as there is no agreed upon standard which can be used to judge whether a particular model can be deemed "explainable" or not. This issue has been accentuated by what the term "explanation" refers to in the context of AI. Is it an approximation of a complex model [1, 24], an assignment of causal responsibility [25] or a set of human intelligible features contributing to a model's prediction [27]? To make matters worse, a variety of circular definitions for "explanation" can be found, alluding to further concepts like "interpretation" and "understanding" which are, in turn, left undefined.

Early work, stemming mostly from philosophy, gravitated around the idea of "explanation" as a perennial carrier of causal information [20, 16]. More recently, we find literature that concentrates on the ethos of XAI with a focus on applications in ML. Most acknowledge the epistemological leniency when talking about "explanations" and terms alike [27, 21, 39]. A recurring motif from this literature is also to define "explanations" or "interpretations" as an agglomeration of different, more specific terms like confidence, transparency and trust [39, 9, 12]. Some work focuses on classifying the requirements that an explainable system should meet [39, 21] or the kind of evaluations through which a model can be deemed explainable [21, 12]. This lack of consensus has resulted in research that, albeit exciting, ends up tackling different problems, making it impossible to list and compare literature solely based on what has been called "explanation" [6, 21].

In order to measure and compare progress in the field of XAI, we argue that a unified foundation is vital, and that such foundation starts with an adequate definition of the field's terminology. In particular, we propose a framework based around atomic notions of "explanation", and "interpretation" in the context of AI (with a special focus on ML and applications in computer vision). We show how further

*Authors contributed equally

concepts mentioned in the literature can be rephrased in relation to our proposed definition of "explanation" and "interpretation", facilitating the comparison of methods claiming to be explainable.

2. Context and definitions

In order to lay down a sound and inclusive foundation, we start by looking at core aspects that most research in XAI share, but also at what makes them incompatible. What counts as an atomic notion and what is treated as a system, has been the prerogative of each scientific contribution. This has been in part acceptable, given how under-specified modern ML tasks are [22, 10]. Starting from an intuitive idea of object classification, we already assume and accept the interaction of signs, objects and interpreters from semiotic theory [35]. These requirements get further reduced to low-level mathematical primitives e.g., the notion of a "chicken" gets represented as a set of points $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$ which get further simplified as tensors of bytes in the range $[0, 255]$.

In turn, XAI is essentially searching for evidence about non-functional requirements of the high-level task (e.g., whether a higher relevance score is being attributed to the area where a target object is) within the low-level primitives such as tensors, probabilities, and model parameters (Figure 1). The way we accept mathematical distributions as evidence for the presence (or absence) of an object in an image follows a well-defined mapping from high-level ideas to low-level primitives. Now, in the absence of well-established mappings between the task's non-functional properties and its corresponding low-level primitives, we are obliged to define one explicitly.

To that end, we propose a framework to establish mappings for non-functional properties, such that existing work is covered by it while providing the required rigor to serve as a vehicle for scientific discussion. We begin by identifying two fundamental characteristics that such a framework must have:

1. **Commensurable:** in order to fairly compare two different methods, common measures need to exist. A common vocabulary is therefore required, upon which these metrics can operate. In particular, what counts as an "explanation" and what is "interpretation" needs to be agreed upon beforehand.

2. **Universal:** a generic workflow has to exist, defining the context that identifies atomic primitives, and operations on those primitives. Comparisons between primitives and operations are possible as long as the context in which they are being compared remains the same.

We begin by looking at general notions and definitions for the terms "explanation" and "interpretation" as a basis

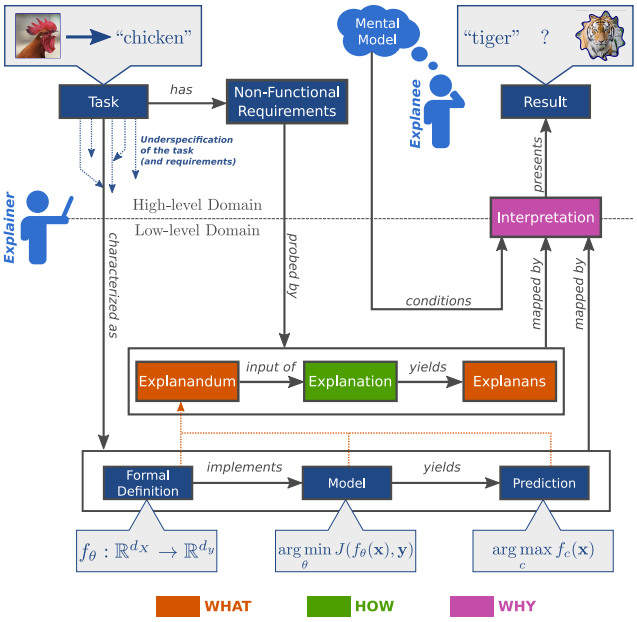

Figure 1. Overview of our proposed framework. A task defined in the high-level domain gets an under-specified characterization, leaving out non-functional requirements. "Explanations" are methods that probe for said requirements. Interpretations are mappings from low- to the high-level domain.

for a refined characterization of said terms in the context of XAI. Moreover, we will identify the minimal requirements for defining a context and therefore, for bounding the scope of explanations and interpretations.

2.1. Towards commensurable explainability: a common vocabulary

Most definitions treat "explanation" and "interpretation" in a similar fashion, sometimes even as synonyms. However, there are subtle but fundamental differences that allow some initial distinction to be drawn between them. First, we need to ask what kind of semantic entity an explanation is: is it an action, an outcome, a process or an object? For most textbook definitions (see Table 2), explanations are seen as statements. In turn, such statements are nothing but descriptions *about* an already existing entity (the *explanandum* or the one which is subject to description). From a functional perspective, an explanation is therefore the process by which an explanandum is described. Finally, to avoid confusions between the process of explaining and its output, we refer to the former as *explanation* and the latter as the *explanans*.

To prevent any kind of circular definition, the explanandum needs to exist axiomatically, thus it has to refer to objects or symbols that are self-evidently true. In other words, we require explanandums to be factual and axiomatic. As we seek explanations for AI models, we find such suitable

Table 1. Definitions of "explanation" and "interpretation" found in XAI literature.

Source	Explanation	Interpretation
Lewis [20]	"someone who is in possession of some information about the causal history of some event (...) tries to convey it to someone else."	—
Josephson & Josephson [16]	"assignment of causal responsibility"	—
Lombrozo [22]	"central to our sense of understanding and the currency in which we exchange beliefs. Explanations often support the broader function of guiding reasoning."	—
Biran & Cotton [7]	—	"the degree to which an observer can understand the cause of a decision"
Lundberg & Lee [24]	"interpretable approximation of the original [complex] model"	—
Montavon et al. [27]	"collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g., classification or regression)"	"mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of"
Dam et al. [9]	"measures the degree to which a human observer can understand the reasons behind a decision (e.g., a prediction) made by the model"	—
Doshi-Velez & Kim [12]	—	"to explain or to present in understandable terms to a human"
Lakkaraju et al. [18]	—	"quantifies how easy it is to understand and reason about the explanation. Depends on the complexity of the explanation"
Vilone & Longo [37]	"the collection of features of an interpretable domain that contributed to produce a prediction for a given item"	"the capacity to provide or bring out the meaning of an abstract concept"
Schmid & Finzel [32]	"in human–human interaction, explanations have the function to make something clear by giving a detailed description, a reason, or justification"	—
Al-Shedivat et al. [1]	"local approximation of a complex model [by another model]"	—

facts in the form of low-level mathematical primitives used to build the models themselves: support vector machines have a decision boundary equation, support vector coordinates, etc. A Neural Network has values for each parameter, equations governing how they connect with each other, a value of the cost function when computed on a particular input, etc. All of these are concrete, undisputed facts (assuming there are no bugs) that are suitable explanandums.

In a simplified, more intuitive form, a first definition of "explanation" can be formulated as follows:

> *An explanation is the process of describing one or more facts.*

The third aspect of an explanation deals with its purpose. Ideally, the output of the explanation (i.e., the explanans) exposes patterns or statistics that were not evident before. For example, overlaying the gradients of an image classifier w.r.t. an input sample $\frac{\partial \mathcal{L}(f(\mathbf{x}),y)}{\partial \mathbf{x}}$ can locate areas that are more sensitive to changes in the input. Once more, resorting to textbook definitions we see that most of them mention "making something understandable, clear, comprehensible" as the goal of an "explanation". In other words, the purpose of an explanation is to enable (human) understanding. Explanations are therefore bound to describe facts in a way that ultimately enable (human) understanding. At this point, we can think of what consumers of explanations (i.e., explainees) need to understand: on the one hand, there are characteristics of the fact being described (e.g., location of the magnitude and sign of a gradient w.r.t. an input sample) which may help one understand what the gradient is[1] On the other hand, there are some characteristics of the described fact in relation to other high-level phenomena e.g., how a high gradient value relates to a latent relevant feature.

[1] Assuming that the explanee's mental model is otherwise equipped with the necessary knowledge to understand this concept.

With this in mind, we arrive at a revised definition for "explanation":

> *An explanation is the process of describing one or more facts, such that it facilitates the understanding of aspects related to said facts (by a human consumer).*

In order to constrain the many ways a description can be interpreted, a contract must be introduced detailing the valid meanings that can be extracted from that description. In other words, there should be an agreement on how to read the symbols of a description e.g., which colors on a heatmap mean high or low values. Such agreement anchors or *assigns* meaning to a primitive entity (in our case, an explanans).

Referring back to textbook definitions for "interpretation", we see how most of them rely on the term "explanation", and thus leading to an inexorable circular definition. The one remaining exception follows the philosophical origin of the word and already lays out the requirements of a contract by defining an "assignment of meaning". In fact, the conveyance of meaning is common to all entries, in one way or another. We go along these lines to sketch the definition of interpretation in XAI as follows:

> *Interpretation is the assignment of meaning (to an explanation).*

For ML, the assigned meaning refers to notions of the high-level task for which the explanans is provided as evidence. An interpretation is therefore bridging the gap between underspecified non-functional requirements of the original task and its representation in formal, low-level primitives (e.g., high Shapley values for pixels on a chicken's beak indicate its correct detection and relation to the class "chicken").

In order for an explanation to fulfill its goal (facilitate understanding), the terminology and symbols known to the explainee i.e., the consumer of the explanation, have to match those used by the explainer i.e., the proponent of the explanation, when assigning meaning to an explanation. In other words, the complexity (otherwise known as parsimony) of the interpretation should not be greater than the explainee's capabilities to fathom its meaning [29, 34]. Sometimes, there are no fundamental reasons to prefer one interpretation over another (e.g., make red represent high values instead of blue or white) as long as one is agreed upon.

Even if the methods to explain (or the agreements to interpret) vary, the process of understanding will always be supported by the same mechanism: description of facts followed by the assignment of meaning for the description itself. While the meaning being assigned can be contested (e.g., as part of the scientific process where one seeks to falsify a statement) the process of assignment cannot.

2.2. Universal context for XAI methods: the *what*, the *how* and the *why*

One of the main aims of equipping ML models with explanation capabilities is to contest previous beliefs w.r.t. a particular prediction by constraining an otherwise undetermined problem [22]. So far, we have already constrained two core definitions in XAI, in an effort to define a unified language that allows us to engage in scientific discussion. Said terminology still needs to fit into a more generic, procedural framework where novel and existing contributions can be placed and thus, compared.

Explanations and interpretations are ultimately aimed at answering questions arising from the underlying process (i.e., the ML model) that issued a prediction. These questions, in their more generic form, correspond to variants of *what*, *how* or *why* queries. Hence, our interest lies in defining a framework that addresses these questions when defining novel explanations and interpretations. In fact, we show that both terms are central for answering all three questions. We describe the scope of the aforementioned questions, and the role that explanations and interpretations both play when answering them.

The *what* defines the domains in which explanations operate. As defined in section 2.1, we find that explanations, as processes, take in elements of a particular *source* domain, and output descriptions that exist in a *target* domain. This notion of "translation" between two domains has been recently promoted [13] although the terminology differs with the one proposed in this work. In mathematics, whenever a function is defined, it is first expressed in terms of the domain and co-domain where the function projects values into e.g., $f : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}$. Similarly, an explanation method should explicitly state *what* is being used as input and *what* is being produced as output. Ideally, the output of the explanation (explanans) should be defined in terms of low-level primitives that are as factually true as the input (e.g., intermediate features, support vectors, gradients). Note that, while inputs in the *source* domain are limited by the model being explained, the *target* domain depends on the explanation method, where options are virtually unlimited[2].

Defining *how* something is being done can be addressed from two levels of abstraction. First, at the system-level, there is the question of *how* the model arrives at a prediction. Second, there is the issue of *how* the explanans is produced. In other words, what are the details behind the process that transforms between domains defined by the *what*, and used by the explanation. The former is precisely the kind of inquiry that XAI methods try to answer and therefore, they cannot be included in our frame-

[2]Practical constraints arise from the cognitive limitations of human minds (e.g., the inability to imagine a tesseract).

Table 2. Definitions of "explanation" and "interpretation" according to various dictionaries. Explanation is referred to as an object while interpretation is more commonly associated to an action. Accessed on 05.01.2021.

Source	Explanation	Interpretation
Merriam-Webster	**act** of making plain or understandable	**action to explain** or tell the meaning of
Cambridge	**the details** or other information that someone gives to make something clear or easy to understand	**an explanation** or opinion of what something means
Oxford	**a statement** or account that makes something clear	**the action of explaining** the meaning of something
Dictionary.com	**statement** made to clarify something and make it understandable	**explain**; action to give or provide the meaning of; explicate; elucidate
Princeton	**a statement** that makes something comprehensible by describing the relevant structure or operation or circumstances etc	**an explanation** of something that is not immediately obvious; a mental representation of the meaning or significance of something
Wikipedia	**a set of statements** usually constructed to describe a set of facts that clarifies the causes, context, and consequences of those facts	A philosophical interpretation is **the assignment of meanings** to various concepts, symbols, or objects under consideration.

work. Instead, this kind of *how* questions will be answered through methods that rely on it[3]. The latter on the other hand, deals with the context of explanations and their answer is essentially reduced to the explanation method itself. Simply put, the answer of *how* the explanandum is mapped to an explanans is defined by the explanation method.

As most XAI literature is devoted to the development of explanation methods (e.g., computation of relevance values [4, 36], mapping of high-level conceptual constructs in intermediate activations [5, 17]), the *how* is almost always thoroughly defined. This has already nurtured insightful debates on whether linear approximations of the original model [24] or even other black-boxes [1] can be considered appropriate explanation methods [31, 33].

Together, the *what* and the *how* already constrain the scope in which explanations are valid. The one remaining aspect is the context in which an explanation is interpreted. An explanans is, by itself, only relevant within the target domain defined by the explanation. It is the interpretation (of the explanans) which will map the low-level representations into a high-level domain where expectations regarding non-functional requirements can be validated or contested. Consider the following interpretation: "the result of an *argmax* operation on the logits of a neural network corresponds to the predicted class". The high-level action of predicting a class has been mapped to an *argmax* operation over a vector. Based on this interpretation of the *argmax* operation, users of neural networks can either accept this notion to gather results, or find shortcomings and propose alternative ways of fulfilling the requirement of a model to predict classes (e.g., through the refinement of the prediction through a hierarchical exclusion graph [11]). The same principle can be applied to explanans and non-functional requirements.

Asking *why* something happens, inescapably relates back to causal effects. While these kind of relationships are among the most useful to discover (as it allows for more control over the effect by adjusting the cause), other non-causal relationships remain valuable in the toolset of explainability. Finding out that a model is unfair, without knowing what the cause of it is, can already be helpful in high-stake scenarios (e.g., by preventing its use). **We say that the *why* relates to the nature of an interpretation.** In short, if the interpretation bares a causal meaning, then the *why* is being defined by the causal link. If the meaning is limited to a correlation, the *why* is left out of the scope of that particular interpretation. Note that, if the explanation method is already based on causal theory [23, 8], the assigned meaning (i.e., the link between the explanans and the high-level (non-)functional requirement) will be more direct and therefore, more likely to withstand scientific scrutiny. The *why* is therefore not mandatory in explanations generated by XAI methods. In any case, the explanation's context can and should be defined, be it causal or based solely on correlations. Proponents of XAI methods are responsible for clearly stating the context in which their explanations can be interpreted.

3. On the completeness of the XAI framework

We show that our proposed framework complies with concepts and desiderata related to explanations and explainable models. In particular, those that have been defined by Alvarez-Melis and Jaakkola [2], and Miller [25]. Furthermore, we discuss evaluation metrics for XAI methods (as defined by Doshi-Velez and Kim [12]) and how they also fit

[3] In fact, all three questions–*what, how, why*–have been proposed as the basis for identifying questions that are answerable through explanations [25].

within our framework.

For Alvarez-Melis and Jaakkola [2], there are three characteristics that explanations need to meet: fidelity, diversity and grounding. Fidelity alludes to the preservation of relevant information; diversity states that only a small number of related and non-overlapping concepts should be used by the explanation, while grounding calls for said concepts to be human-understandable. In our proposed framework, fidelity is guaranteed as explanations are defined as mechanisms to describe or map inputs that are axiomatically valid using a well-defined function. Although defining an absolute number of concepts can be rather subjective, diversity is possible by allowing a designer of explanations to purposely focus on a few concepts on the input or output of the explanation. Grounding is essentially addressed by the interpretation, as it is the mapping from a low-level primitive to a high-level, human-understandable realm of a non-functional requirement. Moreover, explanations require the production of artifacts specifically targeted at enabling human understanding.

Miller [25] has highlighted several aspects of explanations that the XAI community has been mostly unaware of: the social, contrastive and selective nature of explanations, and the irrelevance of probabilities when providing an explanation. When defining the domain and co-domain of an explanation, there are several ways by which contrastive explanations can be offered: either several explanandums can be processed, and their explanans compared, or the explanation method itself expects multiple input pairs (possibly producing output pairs too). Employing causal methods as explanations will inevitably encode counterfactual information (e.g., the result of applying a *do*-operator) enabling a comparison with respect to the observed data. Selective explanations closely relate to the concept of diversity from Alvarez-Melis and Jaakkola [2]. The irrelevance of probabilities mainly states that the best explanation for the average case may not be the best explanation for a particular explainee. While true in some cases, the usefulness of tailored explanations is contingent on the use-case (as providing different explanations for two identical cases may violate fairness constraints). Nonetheless, our framework does not preclude an explanation from doing so if the use-case calls for it. Finally, the social aspect of explanations is reflected by its very definition, as the purpose of explanations is to "enable *human* understanding". Furthermore, interpretations are mappings from low-level to high-level requirements, precisely to make an explanans consumable by humans.

General guidelines for measuring explanations (or better said, their outputs) have been proposed in [12]. These are based on three kinds of evaluations depending on the scope of the application (from generic to specific) and they revolve around the involvement of automatic proxy-tasks, non-expert humans, or domain experts. The use of automated tasks would be amenable primarily to explanations, as they already live in the realm of data structures and mathematical primitives. Evaluations that involve humans will thus be better suited for the interpretations, as mappings to a high-level domain, where the non-functional requirements originate, ultimately affect human understanding (therefore impacting trust, confidence, etc).

A fitting example of quantifiable criteria for explanations can be found in the aforementioned work by Alvarez-Melis and Jaakkola [2]. They propose three, arguably generic characteristics that their proposed explanations should meet, namely explicitness, faithfulness and stability. We see that such properties also refer to aspects defined in our framework: explicitness or "how understandable are the explanations" establishes how clear the interpretation of their provided explanans (in their case, a selection of prototypes describing an expected latent feature) is. Faithfulness or the "true relevance of selected features" is directly addressing the quality of an explanans via counterfactual analysis (i.e., had the selected feature not been there, would the prediction suffer any change?). Finally, stability measures consistency of the explanation for similar inputs. As part of the particular classification problem they work on, said property deals with an expected local Lipschitz continuity which guarantees that similar input samples will yield a similar explanans.

We see how our proposed framework offers a comprehensive language that not only aligns and encompasses previously defined desiderata regarding XAI, but also allows the identification of common ground between a wide array of concepts related to the field.

4. Understanding the state of XAI under our proposed framework

The proposed explanation framework helps establishing unambiguous and commensurable relations across all kinds of XAI contributions. We demonstrate the broad applicability of our framework by re-contextualizing three popular methods, LIME, SHAP and MDNet, showing how and where they compare, while also revealing some of the gaps and complementary properties between them.

Additive feature attribution methods like LIME [30] and SHAP [24] are two frequently used XAI techniques today. Both were introduced as model-agnostic methods, aiming at approximating the behavior of complex models, all without requiring access to their internal variables. On the other hand, MDNet [40] was proposed as an "interpretable medical image network" for diagnosis of bladder cancer through the generation of textual diagnosis along with word-wise attention maps. At first glance, establishing a degree of commensurability (as defined in section 2.1) is not straightforward, especially when dealing with methods that operate on

Table 3. Recontextualization of three popular XAI methods with the help of our proposed framework.

Method		LIME	SHAP	MDNet
What	Source	Model input Model prediction	Model input Model prediction	Model activations
	Target	Linear classifier weights	Linear classifier weights	Word-wise attention matrix
How		1) Input perturbation sampling 2) Inference on class of interest 3) Training of proxy model	1) Input perturbation sampling. 2) Inference on class of interest. 3) Training of unique SHAP solution for proxy model	1) Implicit training of attention module. 2) Generation of attention matrices from activations and LSTM state.
Why	Causal	—	—	—
	Non-causal	Surrogate model weights indicate the local influence of features sampled from a marginal distribution	Approximation of the average contribution of a feature to the prediction.	Approximation of the model's attention during word generation

vastly different domains or whose explanations tackle different sets of non-functional requirements such as MDNet's and LIME's. However, by defining the different elements of all three methods in terms of our proposed framework, it becomes possible to establish comparisons and draw limitations with respect to one another. We discuss said elements in the remaining of this section. For a summary of the ensuing discussion, please refer to Table 3.

4.1. Source domain:

The explanandum of LIME and SHAP comprises the prediction of the target model, as well as an input sample[4]. The target model is treated as a black-box, and typically deals with low-dimensional inputs. Meanwhile, MDNet's textual and visual explanations are derived from an input sample, and from high-dimensional latent variables found within the target model. LIME and SHAP represent the target model's low-level internal processes mainly through counterfactual analysis, while methods exposed to the model's internals can examine the flow, translation and attribution of information as it traverses the model, serving as more direct evidence of a model's decision process. In fact, it has been shown that limiting access to the target model's internal representations makes the creation of adversarial attacks possible, compromising the reliability of the explanation [33]. In other words, it is not enough to rely only on the input domain (of the target model) as the explanandum.

[4]Most *local* explanation methods rely on individual samples from the model's input space as input for the explanation too. Non-local methods like TCAV [17] or S2SNet [28] rely on a group of samples or even a representative sample of the input space.

4.2. Target domain:

The outputs of both LIME and SHAP consist of the weight values from linear surrogate models that are mapped to their corresponding input region (visualized as heatmaps). Instead, MDNet's explanans is composed of a sequence of one-hot encoded words from the language model, where each word is paired with a 14×14 weight matrix that is spatially mapped to the input image (also visualized as a heatmap). The combination of generated text supported by word-wise attribution maps provides a traceable structure from the model's input to the weight matrix. While LIME and SHAP provide explanans that share some of the output space with MDNet's, the relation they bare with any internal representation of the target model is not as traceable.

4.3. How:

LIME and SHAP follow similar explanation strategies characterized by continuously perturbing and evaluating input samples via the target model; results are subsequently approximated through a surrogate linear model. SHAP poses additional constrains to the surrogate's optimization and slightly differs in its sampling scheme. MDNet can be described as a two-stage process: an image feature extraction followed by a concurrent generation of a sensitivity map (originally called "explanation") and textual diagnosis. Both the explanation and classification components of MDNet are trained simultaneously in an end-to-end fashion. The language model is trained using diagnosis texts as the supervisory signal, fulfilling one of its functional requirements. The word-wise attention module learns a set of sensitivity weights with additional constrains on diagnostic labels serving as indirect supervision.

4.4. Why:

Given all structural and low-level components of the explanations under scrutiny, as described by the *what* and the *how*, we now turn to the interpretation of such primitives. The first observation is that none of the aforementioned methods provide interpretations of causal nature. LIME and SHAP cannot provide interpretations grounded in causality due to issues related to off-manifold sampling [14] and the non-excludable inaccuracy when relying on proxy models [31]. Nevertheless, there are non-causal interpretations worth examining.

The weight values derived from LIME's proxy models are interpreted as "influence values". These values convey the relevance of individual input features w.r.t. the prediction of the target model. A small caveat to this interpretation is its limited validity, which applies only to the local neighbourhood around the original input sample. SHAP's explanans, despite baring evident similarities with LIME's, allows the generation of additional global explanans through an accumulation of statistics from multiple local explanations. Furthermore, the notion of feature attribution (i.e., how much the value of each input feature has influenced the model's prediction) inherits the properties of Shapely values. In this case, values are interpreted as payouts to individual features, reflecting how much they contributed to the model's prediction, relative to the remaining input features. MDNet's visual explanations are interpreted as the model's attention w.r.t. a region within the input image while the textual-diagnosis generates a word. In this case, we see how MDNet defines explanans as part of the model's output and not as a separate process, tying predictions and explanans together. The correspondence between feature attribution on the image domain and the sensitivity w.r.t. generated text comes from additional annotations available for the text. The assumption is that latent representations of MDNet align features of image regions and text in a way that is meaningful to humans. The link between the intuitive notion of "attention" and its implementation in one of MDNet's modules, has been recently contested [15, 38], undermining any causal relation drawn from it.

5. Conclusions

To address the growing heterogeneity and lack of agreement on what constitutes an explainable or interpretable model, we introduced a novel theoretical framework to consolidate research and methods developed in the field of explainable AI (XAI). This framework is supported by two fundamental definitions, namely "explanation" and "interpretation". These definitions are further contextualized within a general pipeline that constrains other important primitives like input/output domains, and establishes a divide between low-level mathematical constructs and the high-level, human-understandable realm of (non-)functional requirements.

We show that the proposed framework is compliant with desiderata regarding explanations as defined in previous work. Moreover, existing metrics for XAI methods can be placed within the framework allowing an apples-to-apples comparison between different explanations. Finally, we show a concrete scenario where our framework can help comparing existing XAI methods, showing the extent by which each one addresses different aspects of the explainability pipeline.

5.1. Current trends and practices

We conducted an extensive review of XAI literature, allowing us to identify some current practices in the XAI community, some of which we summarize here in terms of our framework.

Most contributions focus on the development and use of explanation methods while neglecting the role of the interpretation (with notable exceptions like [31, 27]). Explanations, on the other hand, are often defined without explicit definitions of the domain and co-domain (i.e., realm of the explanans and explanandum). More recently, explainable models are trying to include richer objectives through auxiliary tasks: this strategy addresses both the underspecification of the task (conveys some of the sought after non-functional requirements) and expresses some of the invariances that are expected of the task. In domains where AI takes over or assists human-professional workers (e.g., for medical applications), interest often shifts from the prediction to the explanation such that human experts learn and gain new insights about the task.

5.2. Recommendations

Finally, we identify three aspects that future research in XAI should focus on in order to expedite the advancement of the state-of-the-art.

- Before working on the specifics of an explanation, explicitly define the non-functional requirements that a task should fulfill, and that the explanation itself will be probing for. This can be achieved by defining better hypotheses e.g., one that can be falsified [19].
- With each new explanation method a clear, measurable interpretation should be provided.
- Metrics for XAI methods should operate within the same level of abstraction w.r.t. the framework i.e., compare explanans to explanans, explanandum to explanandum, interpretation to interpretation, etc.

Acknowledgments: This work was supported by the BMBF project "ExplAINN" (01IS19074)

References

[1] Maruan Al-Shedivat, Avinava Dubey, and Eric Xing. Contextual explanation networks. *Journal of Machine Learning Research*, 21(194):1–44, 2020. 1, 3, 5

[2] David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7786–7795, 2018. 5, 6

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. 1

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 5

[5] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 5

[6] Adrien Bibal and Benoît Frénay. Interpretability of machine learning models and representations: an introduction. In *ESANN*, 2016. 1

[7] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017. 3

[8] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019. 5

[9] Hoa Khanh Dam, Truyen Tran, and Aditya Ghose. Explainable software analytics. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, pages 53–56, 2018. 1, 3

[10] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020. 2

[11] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *European conference on computer vision*, pages 48–64. Springer, 2014. 5

[12] Finale Doshi-Velez and Been Kim. Considerations for evaluation and generalization in interpretable machine learning. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 3–17. Springer, 2018. 1, 3, 5, 6

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. A disentangling invertible interpretation network for explaining latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9223–9232, 2020. 4

[14] Christopher Frye, Damien de Mijolla, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley-based explainability on the data manifold. *arXiv preprint arXiv:2006.01272*, 2020. 8

[15] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, 2019. 8

[16] John R Josephson and Susan G Josephson. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press, 1996. 1, 3

[17] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 5, 7

[18] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138, 2019. 3

[19] Matthew L. Leavitt and Ari Morcos. Towards falsifiable interpretability research, 2020. 8

[20] David Lewis. Causal explanation, philosophical papers, vol. 2, 1986. 1, 3

[21] Zachary C Lipton. The mythos of model interpretability. *Queue*, 2018. 1

[22] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006. 2, 3, 4

[23] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6979–6987, 2017. 5

[24] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017. 1, 3, 5, 6

[25] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2019. 1, 5, 6

[26] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of inmates running the asylum. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*, 2017. 1

[27] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 2018. 1, 3, 8

[28] Sebastian Palacio, Joachim Folz, Joern Hees, Federico Raue, Damian Borth, and Andreas Dengel. What do deep networks like to see. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 7

[29] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and Interpretable Models*

in Computer Vision and Machine Learning, pages 19–36. Springer, 2018. 4

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016. 6

[31] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 5, 8

[32] Ute Schmid and Bettina Finzel. Mutual explanations for cooperative decision making in medicine. *KI-Künstliche Intelligenz*, pages 1–7, 2020. 3

[33] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. 5, 7

[34] Kacper Sokol and Peter Flach. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 56–67, New York, NY, USA, 2020. Association for Computing Machinery. 4

[35] John F Sowa. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Pub., Reading, MA, 1983. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[37] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020. 3

[38] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, 2019. 8

[39] Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *arXiv preprint arXiv:2004.14545*, 2020. 1

[40] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017. 6