A Theoretical Framework for AI Models Explainability with Application in Biomedicine

Matteo Rizzo
Department of Computer Science
Ca' Foscari University
Venice, Italy
matteo.rizzo@unive.it

Alberto Veneri
Department of Computer Science
Ca' Foscari University & ISTI-CNR
Venice, Italy
alberto.veneri@unive.it

Andrea Albarelli
Department of Computer Science
Ca' Foscari University
Venice, Italy
albarelli@unive.it

Claudio Lucchese
Department of Computer Science
Ca' Foscari University
Venice, Italy
claudio.lucchese@unive.it

Marco Nobile
Department of Computer Science
Ca' Foscari University
Venice, Italy
marco.nobile@unive.it

Cristina Conati
Department of Computer Science
University of British Columbia
Vancouver, Canada
conati@ubc.cs.ca

***Abstract*—EXplainable Artificial Intelligence (XAI) is a vibrant research topic in the artificial intelligence community. It is raising growing interest across methods and domains, especially those involving high stake decision-making, such as the biomedical sector. Much has been written about the subject, yet XAI still lacks shared terminology and a framework capable of providing structural soundness to explanations. In our work, we address these issues by proposing a novel definition of explanation that synthesizes what can be found in the literature. We recognize that explanations are not atomic but the combination of evidence stemming from the model and its input-output mapping, and the human interpretation of this evidence. Furthermore, we fit explanations into the properties of faithfulness (*i.e.*, the explanation is an accurate description of the model's inner workings and decision-making process) and plausibility (*i.e.*, how much the explanation seems convincing to the user). Our theoretical framework simplifies how these properties are operationalized, and it provides new insights into common explanation methods that we analyze as case studies. We also discuss the impact that our framework could have in biomedicine, a very sensitive application domain where XAI can have a central role in generating trust.**

***Index Terms*—explainability, machine learning, biomedicine**

I. Introduction

The advent of Deep-Learning (DL) allowed for raising the accuracy bar of Machine Learning (ML) models for countless tasks and domains. Riding the wave of enthusiasm around such stunning results, DL models have been deployed even in high-stake decision-making environments, not without criticism [1]. These kinds of environments require not only high predictive accuracy but also an *explanation* of why that prediction was made. The need for explanations initiated the discussion around the explainability of DL models, which are known to be "black boxes". In other words, their inner workings are hard for humans to be understood. Who should be accountable for a model-based decision and how a model came to a certain prediction are just some of the questions that drive research on explaining ML models. This is particularly relevant, for instance, in the biomedical field, where human lives are at stake, and understanding the reasoning behind a model's predictions is essential to guarantee safety and avoid costly errors [2]. Furthermore, the ability to explain how a model arrived at a certain conclusion can increase the understanding of the underlying biological mechanisms, enabling more informed support to decision-making for clinicians and researchers. With the first recent attempts of the legislative machinery to make explanations for automatic decisions a user's right [3], the pressure on generating explanations for the ML model's behaviors raised even more. Despite the endeavor of the XAI community to develop either models that are explainable by design [4]–[6] and methods to explain existing black-box models [7]–[9], the way to DL explainability is paved with results that are mostly preliminary and anecdotal in nature (*e.g.*, [10]–[12]). Most notably, it is hard to relate different pieces of research due to a lack of common theoretical grounds capable of supporting and guiding the discussion. In particular, we detect a gap in the literature on foundational issues such as a shared definition of the term "explanation" and the users' role in the design and deployment of explainability for complex ML models. The XAI community suffers from the paucity of common terminology, with only a few attempts of establishing one, focusing more on the distinction among the terms "interpretable", "explainable", and "transparent" rather than the inner structure and meaning of an explanation (*e.g.*, [13]–[15]). Similarly, a lack of an outline of the main theoretical components of the discussion around explainability disperses research, while the current literature finds it hard to provide the involved stakeholders with principled analytical tools to operate on black-box models. This trend has been detected in the delicate field of biomedicine and addressed with context-specific guidelines [16]. In this work, we propose a simple, general, and effective theoretical framework that outlines the core components of the explainability machinery and lays out grounds for a more coherent debate on how

to explain the decisions of ML models. Such a framework is not meant to be set in stone but rather to be used as a common reference among researchers and iteratively improved to fit more and more sophisticated explainability methods and strategies. We hope to provide shared jargon and formal definitions to inform and standardize the discussion around crucial topics of XAI. The core of the proposed theoretical framework is a novel definition of explanation, that draws from existing literature in sociology and philosophy but, at the same time, is easy to operationalize when analyzing a specific approach to explain the predictions made by a model. We conceive an explanation as the interaction of two decoupled components, namely *evidence* and its *interpretation*. Evidence is any sort of information stemming from a ML model, while an interpretation is some semantic meaning that human stakeholders attribute to the evidence to make sense of the model's inner workings. We relate these definitions to crucial properties of explanations, especially *faithfulness* and *plausibility*. Jacovi & Goldberg define faithfulness as "the accurate representation of the causal chain of decision-making processes in a model" [17]. We argue that faithfulness relates in different ways to the elements of the proposed theoretical framework because it assures the interpretation of the evidence is true to how the model actually uses it within its inner reasoning. A property orthogonal to faithfulness is plausibility, namely "the degree to which some explanation is aligned with the user's understanding of the model's decision process" [17]. A follow-up work by Jacovi & Goldberg addresses plausibility as the "property of an explanation of being convincing towards the model prediction, regardless of whether the model was correct or whether the interpretation is faithful" [18]. We relate plausibility to faithfulness and highlight a need for faithfulness to be embedded in explainability methods and strategies, as well as plausibility as an important (yet not indispensable) property of the same. This is particularly true in the biomedical field as a high stake environment, where it is crucial for the explanation to portray the decision-making process of the model accurately. As case studies, we zoom in on the evaluation of faithfulness of some popular DL explanation tools and strategies, such as "attention" [9], [19], Gradient-weighted Class Activation Mapping (Grad-CAM) [20] and SHapley Additive exPlanations (SHAP) [8]. In addition, we look at the faithfulness of models traditionally considered intrinsically interpretable (a notion with distance ourselves to) such as a *linear regressors* and models based on *fuzzy logic*.

II. Designing Explainability

Research in XAI seizes the problem of explaining models for decision-making from multiple perspectives. First of all, we observe that most of the existing literature uses the terms "interpretable" and "explainable" interchangeably, while some have highlighted the semantic nuance that distinguishes the two words [21]. We argue that the term *explainable* (and, by extension *explainability*) is more suited than the term *interpretable* (similarly, by extension, *interpretability*) to describe the property of a model for which effort is made to generate human-understandable clarifications of its decision-making process. The definition of *explanation* is thus crucial and will be discussed extensively in section IV. Our claim follows two rationales: *(i)* the term *interpretation* is used within our proposed framework with a precise meaning that deviates from the current literature and that we deem more accurate (see Section IV-C); *(ii)* we argue against grouping models into *inherently interpretable* and *post-hoc explainable*. Recently, Molnar has defined "intrinsic interpretability" as a property of ML models that are considered fully understandable due to their simple structure (*e.g.*, short decision trees or sparse linear models), while "post hoc explainability" as the need for some models to apply interpretation methods after training [22]. Although principled, we drop this hard distinction by claiming that all models embed a certain degree of explainability. Even though, to the best of our knowledge, no metrics can quantify explainability yet, we can assert that this depends on multiple factors. In particular, a model is as explainable as the explanations that are proposed to the user to justify a certain prediction are effective. Thus, bringing the human into the explainability design loop is key to deploying models that are actually explainable. Consequently, there are models for which it is easier to design explanations (*i.e.*, the so-called *white-box* models, *e.g.*, linear regression, decision trees, rule-based systems, etc.) and models for which the same process is more difficult (*i.e.*, the so-called *black-box* models, *e.g.*, artificial neural networks). The notion of difficulty here is defined by the inner complexity of the model, which relates to the amount of cognitive load the user can sustain. We highlight that the degree of explainability moves along a gradient from black-box to white-box models, without clear-cut thresholds. Nevertheless, in section VI, we show that explanations for both white-box and black-box models fit our proposed framework. Thus they both can be structured homogeneously and more deeply understood by leveraging theoretical tools. Most importantly, we advocate for explainability design as a crucial part of Artificial Intelligence (AI) software development. We endorse Chazette et al., in claiming that explainability should be considered a non-functional requirement in the software design process [23]. Thus explanations for any ML models (and, especially, for DL models) should be accounted for within the initial design of an AI-powered application. Even the most accurate black-box model should not be deployed without an explanation mechanism backing it up, as we cannot be sure whether it learned to discriminate over meaningful or wrong features. A classic example is a dog image classification model learning to detect huskies because of the snowy setting instead of the features of the animal itself, involuntarily deceiving the users [7]. A design-oriented approach to AI development should involve taking humans into the loop, thus fostering a human-centered AI which is more intelligible by design and is expected to increase trust in the end-users [24].

III. Characterising the Inference Process of a Machine Learning Model

In this section, we provide a formal characterization of the inference process of a general ML model, without any constraint on the task. Such a characterization will be used to introduce the terminology which substantiates the main components of our proposed framework of explainability, whose details are provided in section IV. To this end, we define a ML model M as an arbitrarily complex function mapping a *model input* to a *model conclusion* through a sequential composition of *transformation steps*. The whole characterization is exemplified in Figure 1.

A. Elements of the Characterization

Model Input. The model input consists of a set of features, either coming from an observation or synthetically generated.
Model Conclusion. The model conclusion is the final output of the model, which is the outcome of the last link in the chain of transformations over the model input.
Transformation steps. Overall, the decision-making process of M can be represented as a chain of $N > 0$ transformations of the original model input, that are causally related. This causal chain is enforced by model design (*e.g.*, the sequence of layers in the architecture of a neural network or the depth of a decision tree). We call each stage of this causal chain a "transformation step", and we denote it with s_i, for $i \in [1, N]$. The transformation steps advance the computation from the model input to the model output through *transformation functions*.
Transformation functions. Each transformation step s_i relates to a set of n_i "transformation functions" f_{i,m_i}, where $m_i \in [1, n_i]$ indicates one of the possible learnable functions at s_i. Note that, in general, the number of such functions would be infinite, but we discretize it assuming that we are working on a real scenario using some computational machine. The transformation functions are mappings from a feature set $x_{i-1,j}$ to a feature set $x_{i,z}$, with $j \in [1, k_{i-1}]$, $z \in [1, k_i]$ (*i.e.*, the arrows enclosed in the ellipses in Figure 1). The number k_i denotes the cardinality of the set of all possible feature sets generated by all possible learnable transformation functions at step s_i. These transformation functions are generally opaque to the user in the context of the so-called black-box models. At every step in the chain of transformation steps, the model learns one of the possible transformation functions (*i.e.*, the optimal function according to some learning scheme, highlighted with a solid line in Figure 1). That is, the model learns the function $\hat{f}_{i,m_i}$ such that $\hat{f} = \hat{f}_{N,m_N} \circ \ldots \circ \hat{f}_{i,m_i} \circ \ldots \circ \hat{f}_{1,m_1}$ is the overall approximation of the true mapping from the model input to the model conclusion. According to the notation above, we denote the model input as $x_{0,0}$ (or simply x) and the model conclusion as $\hat{y}_{N,j}$, with $j \in [1, k_N]$ (or simply $\hat{y}$).

B. Observations

We asserted that, at each transformation step s_i, the model picks one function $\hat{f}_{i,m_i}$ among n_i such that $\hat{f}_{i,m_i}(x_{i-1}, j) = x_{i,z}$. This raises issues that increase model opacity. At step

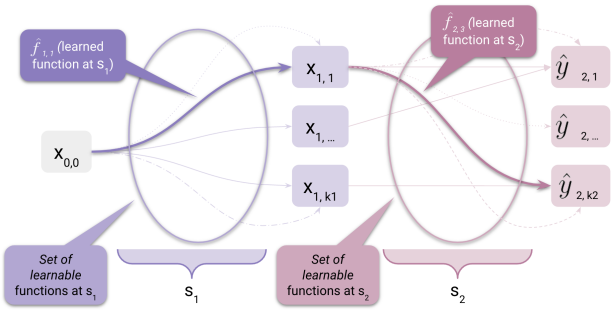

Fig. 1: Example of transformation functions for two steps s_i.

s_i the chosen function $\hat{f}_{i,m_i}$ can map different intermediate transformations $x_{i-1,j}$ of the feature set at the previous transformation step into the same transformation $x_{i,z}$ one step further in the chain. This means that the same outcome in the transformation chain, be it intermediate or conclusive, can be achieved through different rationales, and it could be difficult for a human user to understand which of them is the one the model has actually learned. This can be a result of a high dimensionality of the set of transformation functions, as well as a high complexity of the transformed feature set.

For example, pictures of zebras and salmon can be discriminated on the basis of either their anatomy (i.e., zebras have stripes, salmon have gills) or the environment/habitat (i.e., zebras live in savannas and salmon in rivers). If we consider a relatively complex model such as a Convolutional Neural Network (CNN), where a transformation step coincides with a layer within the network architecture, it is generally difficult to understand which kind of transformation f_{i,m_i} this represents, if any that is human-understandable. Thus, how do we make sense of which of the n_i possible alternative mappings of $x_{i-1,j}$ led to $x_{i,z}$? This remains an open question, with major implications for the discussion around faithfulness, which we will enlarge in the next section.

IV. Defining Explanations

Recent work on ML interpretability produced multiple definitions for the term "explanation". According to Lipton, "explanation refers to numerous ways of exchanging information about a phenomenon, in this case, the functionality of a model or the rationale and criteria for a decision, to different stakeholders" [25]. Similarly, for Guidotti et al. "an explanation is an "interface" between humans and a decision-maker that is at the same time both an accurate proxy of the decision-maker and comprehensible to humans" [26]. Murdoch et al. add to how the explanation is delivered to the user stating that "an explanation is some relevant knowledge extracted from a machine-learning model concerning relationships either contained in data or learned by the model. [...] They can be produced in formats such as visualizations, natural language, or mathematical equations, depending on the context and audience" [27]. On a more general note, Mueller et al. state that "the property of "being an explanation" is not a property of the text, statements, narratives, diagrams,

or other forms of material. It is an interaction of (i) the offered explanation, (ii) the learner's knowledge and beliefs, (iii) the context or situation and its immediate demands, and (iv) the learner's goals or purposes in that context" [28]. Finally, Miller tackles the challenge of defining explanations from a sociological perspective. The author highlights a wide taxonomy of explanations but focuses on those which are an answer to a "why-question" [29].

The definitions mentioned above offer a well-rounded perspective on what constitutes an explanation. However, they fail to highlight its atomic components and to characterize their relationships. We synthesize our proposed definition of explanation based on complementary aspects of the existing definitions. The result is a concise definition that is easy to operationalize for supporting the analysis of multiple approaches to explainability. Our full proposed framework is reported in the scheme in Figure 2, whose components will be discussed in the following sections.

A. Explanation

Given a model M which takes an input x and returns a prediction $\hat{y}$, we define *explanation* the output of an *interpretation function* applied to some *evidence*, providing the answer to a "why question" posed by the user.

B. Evidence

Evidence (e) is whatever kind of objective information stemming from the model we wish to provide an explanation for and that can reveal insights into its inner workings and rationale for prediction (*e.g.*, attention weights, model parameters, gradients, etc.).

1) Evidence Extractor: An *evidence extractor* (ξ) is a method fetching some relevant information about either M, x, $\hat{y}$, or a combination of the three. Then: $e = \xi(x, \hat{y}, M)$. Examples of evidence extractors are, e.g., encoder plus attention layers, gradient back-propagation, and random tree approximation, with the corresponding extracted evidence being attention weights, gradient values, and random tree mimicking the original model. In the peculiar case of a white-box approach, that is, ML models designed to be *easily interpretable* by the user (e.g., linear regression, fuzzy rule-based systems) the extraction of evidence is straightforward since all components of the model directly present a piece of semantic information in a human-comprehensible format.

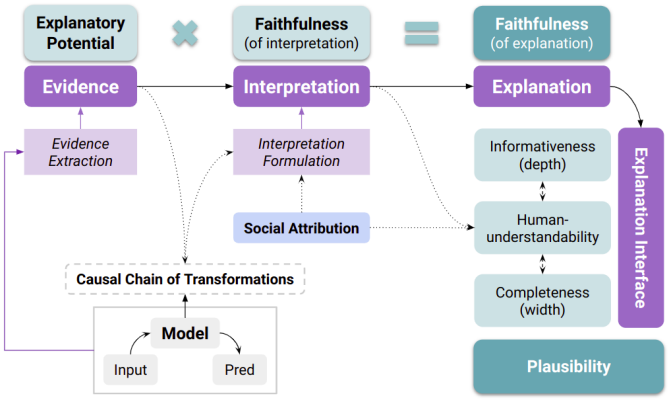

Fig. 2: Overview of the theoretical framework of explainability.

2) Explanatory Potential: We define *explanatory potential* ($\epsilon(e)$) of some evidence as the extent to which the evidence influences the causal chain of transformations steps of a model. Intuitively, the explanatory potential indicates "how much" of a model the selected type of evidence can explain. It can be computed either by counting how many transformation steps are impacted by the evidence (i.e., *breadth*), or how much of each single transformation step is impacted by the evidence (i.e., *depth*).

C. Interpretation

An *interpretation* is a function g associating semantic meaning to some evidence and mapping its instances into explanations either for a given prediction or the whole model. Then an explanation can be defined as either $E = g(e, x, \hat{y}, M)$, or $E = g(e, M)$, respectively.

1) Local vs. Global Interpretations: In accordance with the existing literature, we relate "evidence" and "interpretation" to the concepts of *locality* and *globality*. Both evidence and interpretations can either be local or global. Local evidence (e.g., attention weights, gradient, etc.) relates relevant model information to a particular model input x and corresponding prediction $\hat{y}$. Global evidence (e.g. full set of model parameters) is generally independent of specific inputs and might explain higher level functioning (providing deeper or wider info) of the model or some of its sub-components. Similarly, interpretations can provide either a local or a global semantic of the evidence. A local interpretation of attention could be, *e.g.*, "attention weights are descriptive of input components' importance to model output". On the other hand, a global interpretation of the same evidence may aggregate all the attention weights' heatmaps for a whole dataset and highlight specific patterns. For example, in a dog vs. cat classification problem, a global interpretation of attention may be represented by clusters of similar parts of the animal's body (*e.g.*, groups of ears, tails, etc.) highlighted by the attention activations.

2) Generating Interpretations: Given some evidence involved in one or more steps s_i of M, we guess how this evidence is involved in the opaque input-to-output transformations by formulating an interpretation g of some extent of the decision-making process of the model. At a low level, we generate a candidate g that encapsulates the approximations $f^*_{i,m_i} \tilde{=} \hat{f}_{i,m_i}$ of the behavior of certain functions learned by M at some steps s_i. On an abstract level, interpretations can be seen as hypotheses about the role of evidence in the explanation-generation process. Like a good experimental hypothesis, a good interpretation satisfies two core properties: (i) is testable, and (ii) clearly defines dependent and independent variables. Interpretations can be formulated using different forms of reasoning (*e.g.*, deductive, inductive, abductive, etc.). In particular, the survey on explanations and social sciences

by Miller reports that people usually make assumptions (*i.e.*, in our context, choose an interpretation) via social attribution of intent (to the evidence) [29]. Social attribution is concerned with how people attribute or explain the behavior of others, and not with the real causes of the behavior. Social attribution is generally expressed through folk psychology, which is the attribution of intentional behavior using everyday terms such as beliefs, desires, intentions, emotions, and personality traits. Such concepts may not truly be the cause of the described behavior but are indeed those humans leverage to model and predict each others' behaviors. This may generate misalignment between a hypothesized interpretation of some evidence and its actual role within the inference process of the model. In other words, reasoning on evidence through folk psychology might generate interpretations that are *plausible* but not necessarily *faithful* to the inference process of the model (such terms will be further explored in section V).

D. Explanation Interface

Explanations are meant to be delivered to some target users. We define eXplanation User Interface (XUI) as the format in which some explanation is presented to the end user. This could be, for example, in the form of text, plots, infographics, etc. We argue that an XUI is characterized by three main properties: (i) human understandability, (ii) informativeness, and (iii) completeness. The *human-understandability* is the degree to which users can understand the answer to their "why" question via the XUI. This property depends on user cognition, bias, expertise, goals, etc., and is influenced by the complexity of the selected interpretation function. The *informativeness* (i.e., depth) of an explanation is a measure of the effectiveness of an XUI in answering the why question posed by the user. That is the depth of information for some s_i of great interest in the XUI. The *completeness* (i.e., width) of an explanation is the accuracy of an XUI in describing the overall model's workings, and the degree to which it allows for anticipating predictions. That is the width in terms of the number of s_i the XAI spans. Note that both informativeness and completeness are bound by the explanatory potential of the evidence (*e.g.*, attention weights do not explain the full model, just some transformation steps, while the full set of model parameters does).

V. Concerning Faithfulness and Plausibility

In Section IV-C2 we observed that social attribution is a double-edged sword for the interpretation generation process, as it may incur in propelling plausibility without accounting for faithfulness. This issue was highlighted by Jacovi & Goldberg, who introduced a property of explanations called *aligned faithfulness* [18]. In the words of the authors, an explanation satisfies this property if "it is faithful and aligned to the social attribution of the intent behind the causal chain of decision-making processes". Our proposed framework allows us to go a step forward in the characterization of this property. We note that the property of aligned faithfulness pertains only to interpretations, not evidence. The latter by itself has no inherent meaning, its semantics is defined by some interpretation that may or may not involve social attribution of intent to the causal chain of inference processes.

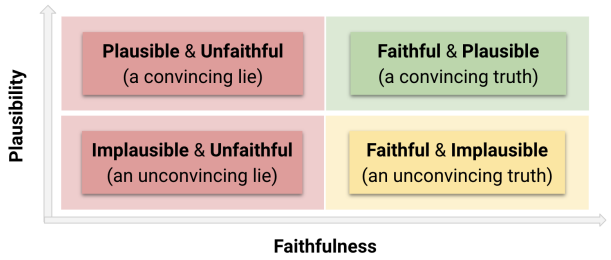

Fig. 3: Overview of the outcome on the user of the interaction between faithfulness and plausibility.

A. Faithfulness

Given an interpretation function g describing some transformation steps s_i within a model M's inference process, we want to be able to prove that g is faithful (at least to some extent) to the actual transformations made by M to an input x to get a prediction $\hat{y}$. Namely, we define the property of *faithfulness of an interpretation* $\phi_i(g, e)$, as the extent to which an interpretation g accurately describes the behavior of some transformation functions f_{i,m_i} that some model learned to map an output $x_{i-1,j}$ at s_{i-1} into $x_{i,z}$ at s_i making use of some instance evidence e. Given some evidence e and its interpretation function g, we say that a related explanation is faithful to some transformation steps if the following conditions hold: (i) the evidence e has explanatory potential $\epsilon_i > 0$, and (ii) the interpretation g has faithfulness $\phi_i > 0$. Then we can define the *faithfulness of an explanation* (Φ) as a function of the faithfulness of the interpretation of each step involved and the related explanatory potential. For example, we could define $\Phi = \sum_i \epsilon_i \phi_i \ \forall i \in I \subseteq [1, N]$ where I is the set of indices of transformation steps s_i that involved the evidence e. Thus the faithfulness of an explanation is the sum of the faithfulness scores of its components, *i.e.*, the faithfulness of the interpretations of the evidence involved in the generation of the explanation. Besides, the related explanatory power weights the faithfulness of each interpretation, following the intuition that evidence with higher ϵ should have a larger impact on the overall faithfulness score of the interpretation.

We can have various measures of faithfulness that are associated with different explanation types, in the same way as we have different metrics to evaluate the ability of a ML model to complete a task. Thus ϕ_i is implicitly bounded.

When designing a faithful explanatory method, we can opt for two approaches. We can achieve faithfulness "structurally" by enforcing this property on pre-selected interpretations in model design (*e.g.*, imposing constraints on transformation steps limiting the range of learnable functions). This direction has been recently explored by Jain et al. [30] and Jacovi & Goldberg [18]. An alternative, naive, strategy is trial-and-error: formulating interpretations and assessing their faithfulness via formal proofs or requirements-based testing using proxy tasks.

While formal proofs are still missing in current literature, a number of tests for faithfulness have been recently proposed [10]–[12], [31].

B. Plausibility

The combined value of the three above-mentioned properties of XUIs (*i.e.*, human understandability, informativeness, and completeness) drives the plausibility of an explanation. More specifically, we define *plausibility* as the degree to which an explanation is aligned with the user's understanding of the model's partial or overall inner workings. Plausibility is a user-dependent property and as such, it is subject to the user's knowledge, bias, etc. Unlike faithfulness, the plausibility of explanations can be assessed via user studies. Note that a plausible explanation is not necessarily faithful, just like a faithful explanation is not necessarily plausible. It is desirable for both properties to be satisfied in the design of some explanation. Interestingly, an unfaithful but plausible explanation may deceive a user into believing that a model behaves according to a rationale when it is actually not the case. This raises ethical concerns around the possibility that poorly designed explanations could spread inaccurate or false knowledge among the end-users. Figure 3 provides a simplified overview of the problem.

VI. Framing Common Explainability Strategies

A. Attention

The introduction of attention mechanisms has been one of the most notable breakthroughs in DL research in recent years. Originally proposed for empowering neural machine translation tasks [9], it is currently employed in many state-of-the-art approaches for numerous cognitive tasks. The chain of transformations in the simplest neural model making use of self-attention is a three-step causal process: (i) encoding, (ii) weight encodings by attention scores, and (iii) decoding into model output. Then we can define the function learned by the model as the composition $\hat{f} = f_{3,m_3} \circ f_{2,m_2} \circ f_{1,m_1}$, where each f_i for $i \in [1,3]$ corresponds to the respective transformation function in the causal chain.

Evidence. For an input x split into t sequentially related tokens, let f_{1,m_1} be an encoder function such that $f_{1,m_1}(x) = \bar{X}$ is the vector of the encoded model input tokens. Then, $f_{2,m_2}(\bar{X}) = \sum_{j=1}^{t} \alpha_j \bar{x}_j$, for all model input tokens $\bar{x}_j \in \bar{X}$, is the linear combination of the encodings weighted by their corresponding attention scores. Then $e_{att} = \{\alpha_j\}_1^t$.

$$e_{att} = \{\alpha_j \mid f_{2,m_2}(\bar{X}) = \sum_{j=1}^{t} \alpha_j \bar{x}_j\} \tag{1}$$

That is, the evidence e_{att} related to a model input is the set of weights α_j produced by the attention layer. The explanatory potential $\epsilon(e_{att})$ is the ratio between the number of parameters involved in the analyzed attention layer with respect to the total number of parameters of the model.

Interpretation. The interpretation of the evidence is a function $g_{att}(e(x, \hat{y}))$ that describes function f_{3,m_3}, *i.e.*, how the weighted encodings are decoded into the model conclusion.

Faithfulness. Note that we do not know the faithful interpretation function, so we hypothesize its behavior by formulating a candidate interpretation, a process that is usually guided by the researcher's intuition. In the case of attention, an interpretation generally shared among researchers is that "the value of each attention weight describes the importance of the corresponding token in the original input to the model output". Unfortunately, albeit plausible, research in this field disproved such an interpretation of attention weights [10]–[12], leaving the role of attention for explainability (if any) still unclear.

B. Grad-CAM

A popular explanation called Grad-CAM [20] presents a method to explain a prediction made by an image classifier using the information encompassed in the back-propagated gradient of a prediction. In short, Grad-CAM uses the information about the gradient computed at the last convolutional layer of a CNN given a certain input x to assign a feature importance score for each input feature.

Evidence. The Grad-CAM evidence-extraction ξ_{grad} method consisted of using the feature activation map of a convolutional layer from a given input x to compute the neurons' importance weights α_i. The explanatory potential $\epsilon(e_{grad})$ is related, as for the attention mechanism, to the number of parameters analyzed w.r.t the total number of parameters of the method.

Interpretation. Grad-CAM claim that the computed neuron's weights α_i corresponds to the part of the input features that influence the final prediction the most.

Faithfulness. The authors measure the faithfulness of the model using image occlusion. That is, they patched some part of the input to the model, and they measured the correlation with the difference in the final output. With this faithfulness metric, a high correlation means a high faithfulness in the explanation.

C. SHAP

Lundberg & Lee in 2017 proposed SHAP [8], a method to assign an importance value to each feature used by an opaque model M to explain a single prediction $\hat{y}$. SHAP has been presented as a generalization of other well-known explanation methods, such as Local Interpretable Model-agnostic Explanations (LIME) [32], DeepLIFT [33], Layer-wise relevance propagation [34], and classic Shapley value estimation [8]. The SHAP values are defined as:

$$h\left(z'\right) = \beta_0 + \sum_{i=1}^{M} \beta_i z_i' \tag{2}$$

where $z_i' \in \{0,1\}^M$ is a simplified version of the input x, M is the number of features used in the explanation, and $\beta_i \in \mathbb{R}$ is a coefficient that represents the effect that the $i-th$ feature has on the output.

Evidence. The only evidence $e_{shape} = \xi_{shap}(M, x)$ used by SHAP is the set of predictions made by the classifier in a neighborhood of x. To compute the explanatory potential $\epsilon(e_{shap})$, we can use the ratio of predictions employed to compute the SHAP values w.r.t the total number of possible samples in the countable (and possibly infinite) neighborhood of x. Thus, the greater the number of predictions we have, the higher the exploratory potential of the method.

Interpretation. The interpretation g_{shap} of the evidence proposed by SHAP is that, given e_{shap}, we can locally reproduce the behavior of a complex unknown model with a simple additive model $h(\cdot)$, and analyzing $h(\cdot)$ we can get a local explanation E_{shap} of the behavior of the initial model. That is, the proposed interpretation of the evidence results from the optimization problem in Equation 2.

Faithfulness. Even though the authors do not present a measure of the faithfulness of the explanation directly, they provide three desirable properties that are *i)* local accuracy, *ii)* missingness, *iii)* consistency. The authors showed that their method is the only one that satisfies all these properties, assessing a requirements-based form of faithfulness as described in § V.

D. Linear regression models

Linear regression models are not an explanation method but are normally considered *intrinsically interpretable*. Following our proposed framework, we claim that defining them, among other models, as *intrinsically interpretable* is inaccurate and often misleading. In fact, the definition of what is simple to be interpreted by humans is not well-defined, and we can enumerate various examples of models that are easy to be interpreted by a practitioner but are almost black-boxed for non-expert users.

A linear regressor $\hat{f}_{lin}(\cdot)$ is typically formulated as:

$$\hat{f}_{lin}(x) = \beta_0 + \sum_{i=1}^{N} \beta_i x_i' \tag{3}$$

where β_i are the weights of the learned features, and N is the feature space dimension.

Evidence. The implicit assumption, claiming that a linear model is intrinsically interpretable, is that the weights $\beta_i, 1 \leq i \leq M$ are a good explanation for the model. Thus $e_{lin} = \{\beta_i\}_1^N$. With a linear model, we have the maximum explanatory potential ξ_{lin} because with e_{lin} we can fully describe the model.

Interpretation. Assuming a normalization of the features, we can say that the higher the value of β_i, the higher the contribution of the feature x_i to the model prediction.

Faithfulness. There are no doubts about the faithfulness of the interpretation of the predictions given the normalization assumption, and in fact, a linear model is normally considered an intrinsically interpretable method. However, in a real scenario, its plausibility to a non-expert user is not guaranteed.

E. Fuzzy models

Fuzzy models, especially in the form of Fuzzy Rule-Based Systems (FRBS), represent effective tools for the modeling of complex systems by using a human-comprehensible linguistic approach. Thanks to these characteristics, they are generally considered white or gray boxes and are often mentioned as good options for interpretable AI [35]. FRBSs perform their inference (i.e., calculate a conclusion) by exploiting a knowledge base composed of linguistic terms and rules. A fuzzy rule is usually expressed as a sentence in the form:

```
IF <antecedent> THEN <consequent>        (4)
```

where `antecedent` is a logic formula created by concatenating clauses like `'X IS a'` with some logical operators, where `T` is a linguistic variable (associated with one input feature) and a is a linguistic term. Thanks to this representation, the antecedent of each rule give an intuitive and human-understandable characterization of some class/group.

Evidence. The rules are good evidence for a large part of the model: they characterize the feature space by using a self-explanatory formalism that can be read and validated, by human operators.

Interpretation. The fuzzy sets, that are used to create the fuzzy terms and evaluate the satisfaction of the antecedents, have self-explanatory interpretations: they define how much a value belongs to a given set by means of membership functions. The fuzzy rules are also self-explanatory. The only part that requires a proper interpretation is the output calculation function. In the case of Sugeno reasoning, such functions can be seen as linear regression models, hence all considerations discussed in Section VI-D remain valid also in the case of fuzzy models.

Faithfulness. Similarly to the case of linear regression models, there are no doubts about the faithfulness of the interpretation of the predictions given a normalization step. However, in the case of special transformations (e.g., log-transformation), some of the intrinsic interpretability might be lost in favor of better fitting to training data [35]. Since it is often the case that features in biomedicine (see, e.g., clinical parameters) follow a log-normal distribution, such transformations are very frequent and delicate.

VII. On the impact on biomedicine

AI models have revolutionized the field of biomedicine, enabling advanced analyses, predictions, and decision-making processes. However, the increasing complexity of AI models, such as deep learning neural networks, has raised concerns regarding their lack of explainability. The present paper provides a common ground on theoretical notions of explainability, as a pre-requisite to a principled discussion and evaluation. The final aim is to catalyze a coherent examination of the impact of AI and explainability by highlighting its significance in facilitating trust, regulatory compliance, and accelerating research and development. Such a topic is particularly well-suited for high-stake environments, such as biomedicine.

The use of black box machine learning models in the biomedical field has been steadily increasing, with many researchers relying on them to make predictions and gain

insights from complex datasets. However, the opacity of these models poses a challenge to their explainability, leading to the question of what criteria should be used to evaluate their explanations [36]. Two key criteria that have been proposed are *faithfulness* and *plausibility*. Faithfulness is important because it allows researchers to understand how the model arrived at its conclusions, identify potential sources of error or bias, and ultimately increase the trustworthiness of the model. In the biomedical field, this is particularly important as the stakes are high when it comes to making accurate predictions about patient outcomes and the decision maker wants to have the most accurate motivation behind the model suggestions. On the other hand, while a faithful explanation may accurately reflect the model's inner workings, it may not be understandable or useful to stakeholders who lack expertise in the technical details of the model. Plausibility is therefore important because all the people involved in the decision-making process gain insights from the model that they can act upon and can communicate these insights to other stakeholders in a way that is meaningful and actionable. Both faithfulness and plausibility are important criteria for evaluating explanations of black-box machine-learning models in the biomedical field. Researchers should strive to balance these criteria to provide the best explanations to the stakeholders involved; this can be achieved by involving them during the entire development of the AI machine models and tools.

Beyond the need for faithfulness and plausibility, the concrete impact of explainability on AI for biomedicine is broad. Explainability plays a crucial role in establishing trust between healthcare professionals and AI systems. In critical biomedical applications, such as disease diagnosis, treatment recommendation, and patient monitoring, transparency in decision-making is essential for clinicians to make informed decisions. By providing interpretable explanations, AI models can help healthcare professionals understand the reasoning behind the model's predictions, leading to increased trust and acceptance [37]. Moreover, regulatory bodies of biomedicine, such as the Food and Drug Administration (FDA) in the United States, require transparency and accountability for AI models used in clinical decision-making. XAI techniques provide an opportunity to meet regulatory standards by enabling model auditing and validation. Through explainability, clinicians and regulatory bodies can assess the risk associated with the deployment of AI models, ensuring patient safety and compliance with ethical guidelines [38]. Finally, explainability can significantly enhance the research and development process in biomedicine. By uncovering the underlying factors and features that contribute to an AI model's decision, researchers can gain valuable insights into disease mechanisms, biomarkers, and potential therapeutic targets.

VIII. Conclusions and Future Work

In this work, we propose a novel theoretical framework that brings order and opportunities for a better design of explanations to the XAI community by introducing formal terminology. The framework allows dissecting explanations into evidence (factual data coming from the model) and interpretation (a hypothesized function that describes how the model uses the evidence). The explanation is the product of the application of the interpretation to the evidence and is presented to the target user via some form of explanation interface. These components allow for designing more principled explanations by defining the atomic components and the properties that enable them. There are three core properties: *(i)* the explanatory potential for the evidence (i.e., how much of the model the evidence can tell about); *(ii)* the faithfulness of the interpretation (i.e., whether the interpretation is true to the decision-making of the model); *(iii)* the plausibility of the explanation interface (i.e., how much the explanation makes sense to the user and is intelligible). We show that the theoretical framework can be applied to explanations coming from a variety of methods, which fit the atomic components we propose. The lesson learned from analyzing explanations over the lent of our proposed framework is that humans (both stakeholder and researcher) should be involved in the design of explainability as soon as possible in the AI-powered software design process, especially in sensitive application domains like biomedicine, where a blind application of black-box approaches hampers the right to an explanation. Involving stakeholders allows for a proper filling of each component in the theoretical framework of explainability, and informs model design. The top-down approach that is established this way propels the human understanding of how AI (and ML in particular) works, possibly fostering user trust in the system. We believe that high stake decision-making domains such as biomedicine would be those which will benefit the most from a more rigorous definition of core concepts of explainability, with opportunities to cement a conscious aid of AI-assisted decisions. Therefore, in future work, we want to apply our theoretical study to a real-case scenario in the biomedical sector and analyze its implementation with the help of human feedback, to better focus on the plausibility analysis of our theoretical framework.

References

[1] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat Mach Intell*, vol. 1, no. 5, pp. 206–215, May 2019.

[2] S. Kundu, "Ai in medicine must be explainable," *Nature Medicine*, vol. 27, no. 8, pp. 1328–1328, Aug 2021.

[3] B. Goodman and S. Flaxman, "European union regulations on algorithmic Decision-Making and a "right to explanation"," *AIMag*, vol. 38, no. 3, pp. 50–57, Oct. 2017.

[4] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[5] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8827–8836.

[6] B.-J. Hou and Z.-H. Zhou, "Learning with interpretable structure from gated RNN," *IEEE Trans Neural Netw Learn Syst*, vol. 31, no. 7, pp. 2267–2279, Jul. 2020.

[7] M. T. Ribeiro, S. Singh, and C. 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, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[10] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of the 2019 Conference of the North*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019.

[11] S. Wiegreffe and Y. 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)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 11–20.

[12] S. Serrano and N. A. Smith, "Is attention interpretable?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2931–2951.

[13] M. Graziani, L. Dutkiewicz, D. Calvaresi, J. Amorim, K. Yordanova, M. Vered, R. Nair, P. Henriques Abreu, T. Blanke, V. Pulignano, J. Prior, L. Lauwaert, W. Reijers, A. Depeursinge, V. Andrearczyk, and H. Müller, "A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences," *Artificial Intelligence Review*, 09 2022.

[14] M.-A. Clinciu and H. Hastie, "A survey of explainable AI terminology," in *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*. Association for Computational Linguistics, 2019, pp. 8–13. [Online]. Available: https://aclanthology.org/W19-8403

[15] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, Oct. 2019, company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1900654116

[16] L. Arbelaez Ossa, G. Starke, G. Lorenzini, J. E. Vogt, D. M. Shaw, and B. S. Elger, "Re-focusing explainability in medicine," *Digit Health*, vol. 8, p. 20552076221074488, Feb. 2022.

[17] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4198–4205. [Online]. Available: https://aclanthology.org/2020.acl-main.386

[18] ——, "Aligning faithful interpretations with their social attribution," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 294–310, Mar. 2021.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[21] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 279–288.

[22] C. Molnar, *Interpretable Machine Learning*, 2nd ed. Independently published (February 28, 2022), 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[23] L. Chazette and K. Schneider, "Explainability as a non-functional requirement: challenges and recommendations," *Requirements Engineering*, vol. 25, no. 4, pp. 493–514, Dec. 2020.

[24] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy AI: From principles to practices," *ACM Comput. Surv.*, Aug. 2022.

[25] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, p. 31–57, jun 2018.

[26] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, aug 2018.

[27] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 44, pp. 22 071–22 080, Oct. 2019.

[28] S. T. Mueller, R. R. Hoffman, W. J. Clancey, A. Emrey, and G. Klein, "Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI," *CoRR*, vol. abs/1902.01876, 2019. [Online]. Available: http://arxiv.org/abs/1902.01876

[29] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019. [Online]. Available: https://doi.org/10.1016/j.artint.2018.07.007

[30] S. Jain, S. Wiegreffe, Y. Pinter, and B. C. Wallace, "Learning to faithfully rationalize by construction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4459–4473. [Online]. Available: https://aclanthology.org/2020.acl-main.409

[31] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., Dec. 2018, pp. 9525–9536.

[32] M. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 97–101. [Online]. Available: https://aclanthology.org/N16-3020

[33] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. Sydney, NSW, Australia: PMLR, 2017, pp. 3145–3153.

[34] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise explanations for Non-Linear classifier decisions by Layer-Wise relevance propagation," *PLoS One*, vol. 10, no. 7, p. e0130140, Jul. 2015.

[35] C. Fuchs, S. Spolaor, U. Kaymak, and M. S. Nobile, "The impact of variable selection and transformation on the interpretability and accuracy of fuzzy models," in *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2022, pp. 1–8.

[36] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, and J. H. Holmes, "A manifesto on explainability for artificial intelligence in medicine," *Artif Intell Med*, vol. 133, p. 102423, Oct. 2022.

[37] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1721–1730. [Online]. Available: https://doi.org/10.1145/2783258.2788613

[38] H. Suresh and J. Guttag, *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3465416.3483305