Proceedings of the 22nd International Conference on Principles of Knowledge Representation and Reasoning
Main Track

Categorical Explaining Functors: Ensuring Coherence in Logical Explanations

Stefano Fioravanti[1,2,3] , **Francesco Giannini**[4] , **Pietro Barbiero**[5] , **Paolo Frazzetto**[1] , **Roberto Confalonieri**[1] , **Fabio Zanasi**[6] , **Nicolò Navarin**[1]
[1]University of Padua, Italy
[2]University of Siena, Italy
[3]Charles University, Czech Republic
[4]Scuola Normale Superiore, Italy
[5]IBM Research Zurich, Switzerland
[6]University College London, United Kingdom
{stefano.fioravanti,paolo.frazzetto,roberto.confalonieri,nicolo.navarin}@unipd.it, francesco.giannini@sns.it, pietro.barbiero@ibm.com, f.zanasi@ucl.ac.uk

Abstract

Post-hoc methods in Explainable AI (XAI) elucidate black-box models by identifying input features critical to the model's decision-making. Recent advancements in these methods have facilitated the generation of logic-based explanations that capture interactions among input features. However, these techniques often encounter critical limitations, notably the inability to ensure logical consistency and fidelity between generated explanations and the model's actual decision-making processes. Such inconsistencies jeopardize the reliability of explanations particularly in high-risk domains. To address this gap, we introduce a novel, theoretically rigorous approach rooted in category theory. Specifically, we propose the concept of an *explaining functor*, which preserves logical entailment structurally between the explanations and the decisions of black-box models. By establishing a categorical framework, our method guarantees the coherence and accuracy of extracted explanations, thus overcoming the common pitfalls associated with heuristic-based explanation methods. We demonstrate the practical efficacy of our theoretical contributions through two synthetic benchmarks that highlight significant reductions in contradictory and unfaithful explanations. Our experiments show how our framework can provide mathematically grounded, compositional, and coherent explanations.

1 Introduction

Explainable Artificial Intelligence (XAI) research seeks to understand otherwise black-box AI systems (Ali et al. 2023). In recent years, XAI methods have been applied in multiple research disciplines (Jiménez-Luna, Grisoni, and Schneider 2020; Davies et al. 2021; Giannini et al. 2023; Keskin et al. 2023; Schmidt and Lipson 2009). Post-hoc techniques (Palacio et al. 2021; Ribeiro, Singh, and Guestrin 2016; Ribeiro, Singh, and Guestrin 2018; Ghorbani et al. 2019; Dominici et al. 2024) represent some of the earliest and most common methods as they enable the extraction of explanations from black-box models. Post-hoc explanations typically aim to rank the input features of an opaque model from the most to the least relevant for the model's reasoning. Some of the most advanced post-hoc approaches go beyond raw feature attribution by considering the mutual interaction between input features in the form of logic rules over human-understandable high-level units of information (Ghorbani et al. 2019; Ciravegna et al. 2020), thus enhancing the interpretability of the decision process of the model (Barbiero et al. 2022; Ciravegna et al. 2023; Guidotti et al. 2024; Bobillo et al. 2024). These approaches attempt to reverse-engineer the model's reasoning, translating opaque continuous neural computations into human-understandable Boolean logic formulas, generally through a process of discretization. However, the process of extracting discrete rules from continuous models often fails to guarantee fidelity and logical consistency of explanations.

Limitations of current approaches. To show a critical limitation of approaches based on Boolean logic explanation of continuous functions, as an example let us consider the continuous function $f(x, y) = \min\{1, x + y\}$ defined on the real-unit square, which corresponds to the t-conorm in the Łukasiewicz fuzzy logic (Hájek 2013). A common choice to build a Boolean explanation of a fuzzy function (Jain et al. 2022; Barbiero et al. 2022; Ciravegna et al. 2023) consists in discretizing the input/output representations of the original continuous function, e.g. by thresholding the values at 0.5. However, such a naive discretization approach may directly introduce a discrepancy between the original function and its explanation. For simplicity, in the following we denote by $\bar{x} \in \{0, 1\}$ the discretization of $x \in [0, 1]$ w.r.t. 0.5, i.e. $\bar{x} = 1$ if $x \geq 0.5$, and $\bar{x} = 0$ otherwise. For instance, let us consider the samples $(x_1 = 0.2, y_1 = 0.2)$ and $(x_2 = 0.2, y_2 = 0.4)$, and assume the function f is modeling a classifier in a binary-classification problem. The model classifies the two samples as belonging to the *negative* and *positive* class respectively, indeed $f(x_1, y_1) = 0.4$ and $f(x_2, y_2) = 0.6$ (equivalently expressed by the facts $\overline{f(x_1, y_1)} = 0$ and $\overline{f(x_2, y_2)} = 1$). However, the explanations obtained by thresholding the input components of these opposite samples would be the same, indeed $(\overline{x_1}, \overline{y_1}) = (0, 0) = (\overline{x_2}, \overline{y_2})$. Specifically, the explaining argument would sound as odd as: the sample (x_1, y_1) is classified as *negative* since x_1 and y_1 are negative (below threshold), whereas the sample (x_2, y_2) is classified as *positive* since x_2 and y_2 are negative (below threshold). We refer to this kind of explanations as *inconsistent* w.r.t. the original model, as the same rule $\neg x \wedge \neg y$ explains opposite class predictions $\overline{f(x_1, y_1)} = 0 \neq \overline{f(x_2, y_2)} = 1$. This kind of rule fails to faithfully reflect the model's behavior, leading to misleading or unreliable explanations. This limitation clearly undermines the applicability of these approaches in critical and high-risk decision-making domains since they do not provide explanations that users can trust.

Furthermore, in deep learning the models are composed of multiple layers or modules. Thus, some explanation methods are designed to exploit such compositionality, e.g., Concept Bottleneck Models (CBMs) (Koh et al. 2020; Poeta et al. 2023). However, a well-fitting explanation of deep neural networks should guarantee that composing the different levels of explanations produces an overall explanation that is consistent with respect to explaining the model as a whole. This is often not the case with existing approaches (Ciravegna et al. 2023; Dominici et al. 2025; Debot et al. 2025). These discrepancies imply that current post-hoc explainability methods do not respect functional composition, meaning that explanations derived from smaller, interpretable steps may not be reflected by the behavior of the composed function. These practical observations motivate our work to provide a mathematically solid theory of XAI methods that ensures coherence over composition and fidelity of the extracted explanations, moving beyond the limitations of current heuristic-based approaches.

Our approach. To achieve this goal, we rely on category theory as a mathematical field designed to deal with *processes* and their composition, thereby providing a robust mathematical framework to study compositionality between explainers and classifiers. Indeed, category theory has been widely applied in theoretical computer science (Abramsky and Coecke 2004; Stein and Staton 2021), and more recently in AI (Shiebler, Gavranović, and Wilson 2021; Cruttwell et al. 2022; Giannini et al. 2024; Colombini et al. 2025). In this paper, we take a new perspective introducing the notion of *explaining functor*, a mapping between functions representing concept-based neural networks into Boolean explanations. By defining objects as fuzzy spaces and morphisms as explanatory transformations, our functor structurally preserves logical entailment, thus mitigating the prevalent issue of contradictory or unfaithful explanations. This approach ensures that explanations are not only human-readable, but also logically sound and consistent with the model's internal working. As a proof-of-concept, we show how the proposed theoretical constructions can be used experimentally on synthetic benchmarks based on logic operators. To this end, we use the interpretable deep learning model of Logic Explained Networks (LENs) (Ciravegna et al. 2023), which can provide explanations in terms of logical formulas involving input predicates. Nonetheless, the authors show that LENs can fail to provide coherent explanations. The experiments demonstrate how the theoretical framework can be used in practice to mitigate this issue in this kind of interpretable models.

Contributions. The contributions of this paper can be summarized as follows:

1. We introduce and characterize categories of functions designed explicitly to ensure that their Boolean explanations remain logically consistent and composable.
2. We formally define and analyze categorical functors that systematically map concept-based fuzzy functions to logic formulas, preserving their structural coherence and interpretability.
3. We validate our theoretical framework with practical examples, demonstrating that the proposed categorical functors reliably produce coherent and faithful explanations.

2 Related Work

Categorical foundations in (X)AI. The category theoretic perspective on machine learning has been increasingly adopted in recent years, see Shiebler, Gavranović, and Wilson (2021) for a survey. Particularly close in spirit to our work are approaches that use functoriality to enable compositional analysis of machine learning procedures (Giannini et al. 2024). These works focus on backpropagation and gradient-based learning. Our approach leverages similar categorical methods, but for a different purpose, as it focuses on (logic-based) explainability. Two recent categorical frameworks related with interpretability are Tull et al. (2024) and Rodatz et al. (2024) (but see also Duneau et al. (2024) for compositional interpretability in the context of quantum processes). The work of Rodatz et al. (2024) models a different interpretation notion from ours, whereas Tull et al. (2024) are concerned with defining elementary structures for interpretation at a higher level of abstraction. We expect the maturation of our framework to lead to a convergence of these approaches.

Rule-based explanations. The framework we propose aligns with and extends previous approaches to interpretable machine learning by ensuring logical consistency across explanations at multiple levels of abstraction. For example, LORE (Guidotti et al. 2024) constructs local interpretable predictors and derives decision and counterfactual rules, offering instance-specific explanations. Letham et al. (2015) introduce a generative model called Bayesian Rule Lists that yields a posterior distribution over possible decision lists. ANCHORS (Ribeiro, Singh, and Guestrin 2018) provides local explanations based on if-then logical rules. Our explaining functor differs from these approaches by treating explanations as morphisms that preserve logical entailment across different levels of abstraction. For instance, we may ensure that local explanations remain logically coherent within a broader explanatory structure. LENs (Ciravegna et al. 2023) are one of the most related approaches to our framework, as they are interpretable-by-design neural networks that can be easily converted into Boolean logic formulas. However, LENs' explanations were not perfectly coherent, yielding a limited fidelity of the explanations w.r.t. the original networks. In this sense, the explaining functor yields a practical advantage, maintaining logical consistency while mapping between explanatory levels.

Formal approaches to XAI. The intrinsic limitations of non-formal approaches to XAI motivated the development of formal approaches to XAI in recent years (Marques-Silva and Ignatiev 2022; Arenas et al. 2021; Audemard et al. 2021). In contrast with non-formal approaches to XAI, formal XAI rigorously defines explanations that are sound w.r.t. the model being explained. Nonetheless, these approaches suffer from a number of limitations (Marques-Silva 2024), among them scalability, the size of formal explanations which can be too large, or from the underlying assumption made that any point in the feature space is possible. Finally, they can be applied only to classification problems. The present approach mitigates some of these limitations; the categorical approach provides generalizability to, in principle, any prediction task, and it can be used to provide logically coherent explanations of continuous functions, e.g. implemented by neural networks, as we shall see.

3 Background

3.1 Logic Explained Networks

Logic Explained Networks (LENs) (Ciravegna et al. 2023) are interpretable neural network models designed to provide human-understandable explanations for their decisions using First-Order Logic (FOL) formulas. The main idea behind LENs is to combine deep learning's predictive power with the interpretability of symbolic logic, thereby addressing the typical black-box limitations associated with neural networks.

Formally, a LEN is defined as a function:

$$f : C \to E$$

where $C = [0,1]^k$ is the input concept space representing human-understandable predicate values, and $E = [0,1]^r$ is the output concept space. Each dimension in these spaces corresponds to interpretable concepts. Given an input $c \in C$, the network computes an output $e \in E$, where each dimension e_i is associated with a logical explanation in FOL. More concretely, the output e_i can be explained by a logic rule φ_i, typically provided in disjunctive normal form:

$$\varphi_i(c) = \bigvee_j \left(\bigwedge_l \tilde{c}_{jl} \right), \quad \tilde{c}_{jl} \in \{c_l, \neg c_l\}$$

where each conjunction $\bigwedge_l \tilde{c}_{jl}$ represents conditions under which the output concept e_i is activated. The final FOL formula linking inputs to the predictions of the network is thus:

$$\rho_i = \forall c \in C : e_i(c) \leftrightarrow \varphi_i(c)$$

LENs can be used as explainable-by-design classifiers or as post-hoc explainers of pretrained black-box classifiers.

3.2 Preliminaries of Category Theory

For the basics of category theory we refer the reader to (Mac Lane 1978). A *category* $\mathcal{C}$ consists of a collection of *objects* $\mathcal{C}^o$ and a collection of *morphisms* $\mathcal{C}^h$ between pairs of objects. We write $f\colon X \to Y$ for a morphism from object X to object Y, and $\mathrm{Hom}(X,Y)$ for the set of such morphisms. For all morphisms $f\colon X \to Y$ and $g\colon Y \to Z$ and for all $X \in \mathcal{C}^o$, we require:

1. the existence of a *composite* morphism $g \circ f\colon X \to Z$, with composition being associative;
2. the existence of an *identity* morphism $\mathrm{Id}_X \in \mathrm{Hom}(X,X)$ satisfying
$$f \circ \mathrm{Id}_X = f = \mathrm{Id}_Y \circ f$$

Compared to sets, categories are suitable for studying groups of objects that maintain a particular structure, along with the transformations that preserve this structure. A basic example of this notion is the category whose objects are vector spaces and morphisms are linear maps. In general, any class of general algebras of the same type with their relative homomorphism forms a category.

As examples of two categories central to our development, we mention:

1. the *category of Boolean Functions* $\mathcal{B} = (\mathcal{B}^o, \mathcal{B}^h)$ where $\mathcal{B}^o = \{\{0,1\}^n \mid n \in \mathbb{N}\}$ with $\mathrm{Hom}_{\mathcal{B}}(X,Y) = \{f\colon X \to Y \mid f \text{ is a function}\}$, for all $X, Y \in \mathcal{B}^o$;
2. the *category of Fuzzy Functions* $\mathcal{F} = (\mathcal{F}^o, \mathcal{F}^h)$ where $\mathcal{F}^o = \{[0,1]^n \mid n \in \mathbb{N}\}$ with, $\mathrm{Hom}_{\mathcal{F}}(X,Y) = \{f\colon X \to Y \mid f \text{ is a function}\}$, for all $X, Y \in \mathcal{F}^o$.

Category theory allows us to model not only maps between objects within a given category, but also transformations involving categories themselves. A *functor* models the notion of structure-preserving map between categories. In particular, given two categories $\mathcal{C} = (\mathcal{C}^o, \mathcal{C}^h)$ and $\mathcal{D} = (\mathcal{D}^o, \mathcal{D}^h)$, a *functor* $F\colon \mathcal{C} \to \mathcal{D}$ is a map assigning to each object $X \in \mathcal{C}^o$ an object $F(X) \in \mathcal{D}^o$, and to each $f\colon X \to Y$ in $\mathcal{C}^h$ a morphism $F(f)\colon F(X) \to F(Y)$ in $\mathcal{D}^h$, satisfying $F(Id_X) = Id_{F(X)}$, for all $X \in \mathcal{B}^o$, and $F(f \circ g) = F(f) \circ F(g)$.

Preservation of composition is a particularly relevant property for practical purposes: it amounts to saying that we may compute the value of F on a morphism h entirely in terms of the values of F on $h_1, h_2, \ldots, h_k$, where $h_1 \circ h_2 \circ \cdots \circ h_k = h$ is some decomposition of h into "simpler" morphisms. This property is known as *compositionality* and it enables scaling the complexity of the transformation captured by F.

4 Explaining Functor from δ-COH to Boolean Functions

In this section we introduce a special class of fuzzy functions, the δ-COH functions, which is composed of functions with coherent Boolean explanations (Boolean logic rules). Specifically, when the inputs and outputs are discretized according to a projection $\delta\colon [0,1] \to \{0,1\}$, they produce a well-defined function.

We prove that this class can be equipped with a categorical structure (Theorem 1), being closed under composition. Furthermore, we define a functor that maps this category to the category of Boolean sets with Boolean functions (Theorem 2), thereby providing a rigorous mathematical framework for their interpretation and transformation.

4.1 Category of δ-Coherent Functions

Definition 1. Let $\delta\colon [0,1] \to S$, where $S \subseteq [0,1]$. Then δ is called a *projection* over S if $\delta \circ \delta = \delta$, i.e.,

$$\delta(x) = x \text{ for all } x \in S \quad \text{(idempotency).}$$

The concept of projection onto smaller spaces embodies a fundamental aspect of explainability: the transformation of complex expressions defined over large domains into simpler representations within a more restricted set. It is important to remark that the theory developed in the following sections can be completely generalized outside $[0,1]$.

From now on, we will use the same notation for δ and for all its component-wise applications, i.e., $\delta(x) = [\delta(x_1), \ldots, \delta(x_n)]$, for all $x \in [0,1]^n$. We use this convention since it makes the mathematical notions and proofs more compact and easier to understand.

Definition 2. Let δ be a projection. A function $f\colon [0,1]^n \to [0,1]^m$ is said to be *δ-coherent on* $x \in [0,1]^n$, if $\delta \circ f(x) = \delta \circ f \circ \delta(x)$, i.e., if

$$\delta(f(x)) = \delta(f(\delta(x))).$$

We denote by $\delta\text{-COH}(f) = \{x \in [0,1]^n \mid f \text{ is } \delta\text{-coherent on } x\}$. f is said *δ-coherent* if $\delta\text{-COH}(f) = [0,1]^n$.

We notice that if a function $f\colon [0,1]^n \to [0,1]^m$ is δ-coherent on a certain $x \in [0,1]^n$, then it is also δ-coherent with every $y \in [0,1]^n$ such that $\delta(x) = \delta(y)$.

Given a fixed $\delta\colon [0,1] \to S$, we will often denote $\delta(x)$ by x_δ. Furthermore, for a given $f\colon [0,1]^n \to [0,1]^m$ we denote by f^δ the function $f^\delta\colon x \mapsto \delta(f(x))$.
An example of Definition 2 is given by Boolean threshold functions. In particular, let $\alpha \in [0,1]$, then we call $\delta_\alpha\colon [0,1] \to \{0,1\}$ *α-booleanization* if $\delta_\alpha(x) = 1$ if and only if $x \geq \alpha$, and 0 otherwise, for all $x \in [0,1]$. Another example of Definition 2 is the following:

Lemma 1. *Every constant fuzzy function f is δ-coherent, for any projection δ.*

Proof. Let $f\colon [0,1]^n \to [0,1]^m$ be a constant function with value $c \in [0,1]^m$. Then, $\delta \circ f(x) = \delta(c) = \delta \circ f \circ \delta(x)$. $\square$

We can see that we can easily build examples of functions satisfying or not Definition 2. Let us consider the α-booleanization function $\delta_{0.5}$ where $\alpha = 0.5$. Then trivially the minimum t-norm $\min(x,y)$ is $\delta_{0.5}$-coherent, indeed $\delta_{0.5}(\min(x,y)) = \delta_{0.5}(\min(\delta_{0.5}(x), \delta_{0.5}(y)))$. However, this is not the case for the Łukasiewicz t-norm $x \otimes y = \max(0, x+y-1)$, for example, by taking $x = y = 0.6$, $\delta(x \otimes y) \neq \delta(\delta(x) \otimes \delta(y))$. Indeed, we have $\delta_{0.5}(0.6, 0.6) = (1,1) = \delta_{0.5}(1,1)$ while $\delta_{0.5}(0.6 \otimes 0.6) = \delta_{0.5}(0.2) = 0 \neq 1 = \delta_{0.5}(1) = \delta_{0.5}(1 \otimes 1)$.

Example 1. Consider a fuzzy classifier $f\colon [0,1]^n \to \{0,1\}$ and consider a projection δ. Whenever there exist two points $x, y \in [0,1]^n$ mapped by f in different classes such that $\delta(x) = \delta(y)$, then f is not δ-coherent. Generalizing δ projections over $\mathbb{R}$, consider a binary classifier $f\colon \mathbb{R}^n \to \{0,1\}$ and a projection $\delta\colon \mathbb{R} \to \{0,1\}$. Analogously to the previous case, whenever there exist two points $x, y \in [0,1]^n$ mapped by f in different classes such that $\delta(x) = \delta(y)$, then f is not δ-coherent.

Example 2. Consider a binary classifier $f\colon [0,1]^2 \to \{0,1\}$ such that

$$f(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 < 0.4 \wedge x_2 > 0.6 \\ 0 & \text{otherwise} \end{cases}$$

and consider the threshold function $\delta_{0.5}$. Assume that an explainer provides explanations through Boolean formulas over the discretized inputs (e.g., the explanation of a sample such that $f(x_1, x_2) = 1$ would be $\neg x_1 \wedge x_2$ since $\delta_{0.5}(x_1) = 0 \wedge \delta_{0.5}(x_2) = 1$). Then, we see that f is not $\delta_{0.5}$-coherent since $f(0.3, 0.7) = 1$ and $f(0.4, 0.55) = 0$, for example. These images of f show that the explainer cannot reach 100% accuracy. This is because f is not $\delta_{0.5}$-coherent.

The previous examples illustrate generic and specific non-δ-coherent functions that can be found in any scenario. Moreover, they demonstrate that the lack of δ-coherency hinders the possibility of self-explaining a function by projecting its input through a fixed δ projection, since the set of input-output pairs obtained through projection does not determine a well-defined function.

Theorem 1. *Let $\delta\colon [0,1] \to S \subseteq [0,1]$. The pair $\mathcal{C}_\delta = (\mathcal{C}_\delta^o, \mathcal{C}_\delta^h)$ where $\mathcal{C}_\delta^o = \{[0,1]^n \mid n \in \mathbb{N}\}$ and for all $X, Y \in \mathcal{C}_\delta^o$, $\mathrm{Hom}_{\mathcal{C}_\delta}(X,Y) = \{f\colon X \to Y \mid f \text{ is } \delta-\text{coherent}\}$ is a category called the category of δ-coherent functions δ-COH.*

Proof. For every $X = [0,1]^n \in \mathcal{C}_\delta^o$, the identity morphism 1_X is δ-coherent by definition, and thus $1_X \in \mathrm{Hom}_{\mathcal{C}_\delta}(X,X)$. Assuming $f \in \mathrm{Hom}_{\mathcal{C}_\delta}([0,1]^n, [0,1]^m)$, $g \in \mathrm{Hom}_{\mathcal{C}_\delta}([0,1]^m, [0,1]^k)$, we prove that $g \circ f \in \mathrm{Hom}_{\mathcal{C}_\delta}([0,1]^n, [0,1]^k)$. For this purpose, it suffices to show that $g \circ f$ is δ-coherent. We can see that $\delta \circ f \circ \delta = \delta \circ f$ and $\delta \circ g \circ \delta = \delta \circ g$ since f and g are δ-coherent and, in turn,

$$\delta \circ f \circ g \circ \delta = \delta \circ f \circ \delta \circ g \circ \delta = \delta \circ f \circ \delta \circ g = \delta \circ f \circ g$$

and the claim holds. $\square$

Remark 1. Notice that the choice of any certain projection δ determines a set of different δ-coherent functions and, in turn, a different category.

4.2 Functor from δ-Coherent Functions to Boolean Functions

For each Boolean function $f : \{0,1\}^n \to \{0,1\}^m$, we can consider m Boolean formulas φ_{f_i}, whose truth values are defined by f_i, where $i \in \{1, \ldots, m\}$. As Boolean functions correspond to logic formulas, they are interpretable by design. Now we will see how we associate to each δ-coherent function its Boolean explanation by using a categorical functor that we call the *explaining functor*. In particular, we define a functor from $\mathcal{C}_\delta$ to $\mathcal{B}$ for any given projection $\delta\colon [0,1] \to \{0,1\}$.

Definition 3. Let δ be a projection and $F_\delta : \mathcal{C}_\delta \to \mathcal{B}$ be defined by:

1. for every $X = [0,1]^n \in \mathcal{C}_\delta^o$, $F_\delta(X) = \{0,1\}^n \in \mathcal{B}^o$;
2. for all $f \in \mathrm{Hom}_{\mathcal{C}_\delta}([0,1]^n, [0,1]^m)$, we define $F_\delta(f) \in \mathrm{Hom}_{\mathcal{B}}(\{0,1\}^n, \{0,1\}^m)$, such that $F_\delta(f) = \delta \circ f$.

In the following lemma and theorem we denote $F_\delta(f)$ by f^δ and call f^δ the *δ-function* of f.

Lemma 2. *Let δ be a projection and let $f \in Hom_{\mathcal{C}_\delta}([0,1]^n, [0,1]^m)$. Then. for all $x \in [0,1]^n$, we have:*

$$\delta \circ f = \delta \circ f \circ \delta = f^\delta \circ \delta.$$

Proof. It follows from the fact that f is δ-coherent and $\delta \circ \delta = \delta$. □

Theorem 2. *Let δ be a projection. Then $F_\delta\colon \mathcal{C}_\delta \to \mathcal{B}$ (Definition 3) is a functor.*

Proof. First of all, we have to prove that for every object $[0,1]^n = A \in \mathcal{C}_\delta^o$, $F_\delta(\mathrm{Id}_A) = \mathrm{Id}_A^\delta = \mathrm{Id}_{F_\delta(A)}$. Indeed, $\mathrm{Id}_A^\delta = \delta \circ \mathrm{Id}_A = \delta_{|A} = \mathrm{Id}_{F_\delta(A)}$[1]. Furthermore, let $f \in \mathrm{Hom}_{\mathcal{C}_\delta}([0,1]^n, [0,1]^m)$ and $g \in \mathrm{Hom}_{\mathcal{C}_\delta}([0,1]^m, [0,1]^k)$ we have to prove that $(g \circ f)^\delta = g^\delta \circ f^\delta$. By Lemma 2 we have $\delta \circ g = g^\delta \circ \delta$, and hence

$$(g \circ f)^\delta = \delta \circ g \circ f = \delta \circ g \circ \delta \circ f = g^\delta \circ \delta \circ f = g^\delta \circ f^\delta.$$

□

5 Extending the Explaining Functor

This section provides a method for extending the categorical structure that equips δ-coherent fuzzy functions to all fuzzy functions. As we saw, not all fuzzy functions are δ-coherent; however, this assumption is fundamental to get a functor to the category $\mathcal{B}$. For instance, removing the assumption of δ-coherency, it is easy to produce a function not satisfying Theorem 2.

Example 3. Let $\delta_\alpha\colon [0,1] \to \{0,1\}$ be an α-booleanization with $\alpha = 0.5$, let $g\colon [0,1] \to [0,1]$ with $g(0) = 0$ and $g(x) = 1$ for all $x > 0$, and let $f\colon [0,1] \to [0,1]$ such that $f(x) = 0.2$ if $x < 0.5$ and $f(x) = 1$ otherwise. Then:

$$(g \circ f)^{\delta_\alpha}(0) = \delta_\alpha(g(f(0))) = \delta_\alpha(g(0.2)) = 1 \neq$$
$$0 = \delta_\alpha(0) = \delta_\alpha(g(\delta_\alpha(f(0)))) = g^{\delta_\alpha} \circ f^{\delta_\alpha}(0).$$

In particular, we can see that $g^{\delta_\alpha} \circ f^{\delta_\alpha} = \mathrm{Id}_{\{0,1\}}$ while $(g \circ f)^{\delta_\alpha}$ is the constant 1 function. Note that g is not δ_α-coherent since $\delta(g(0.2)) \neq \delta(g(\delta(0.2)))$.

[1]Here we denote by $\delta_{|A}$ the restriction of δ over A

This example shows how easy it is to find misleading explanations as the composition of two different ones. In fact, let us consider $\delta_\alpha \circ f$ and $\delta_\alpha \circ g$ as fuzzy classifiers. A possible explanation for the classifications could be constituted by some formula of the δ_α projection of the inputs, seen as concepts expressed by Boolean vectors. This example shows that the failure of δ_α-coherency prevents this possibility, since composite functions may lead to contradictory explanations. For this reason, we are interested in defining a mapping that associates to each fuzzy function a δ-coherent function, by preserving the functor compositionality as obtained above.

Transforming a non-δ-coherent to a δ-coherent function may require a non-negligible modification of the original function, which can be carried out in many ways. In this paper, we focus on two main kinds of approaches to produce a δ-coherent from a non-δ-coherent function preserving the points where the latter is coherent: 1) "correct" the input, by expanding the domain to disambiguate input points witnessing a failure of δ-coherency with respect to their δ-functions, 2) "correct" the output on inputs that witness a failure of δ-coherency.

5.1 Domain Extension

The next theorem provides a criterion to associate a coherent function to a non-coherent one via domain extension. Note that we only need to extend the domain for components that are non-δ-coherent.

Theorem 3. *Let δ be a projection, let $f : [0,1]^n \to [0,1]^m$, and let $p \leq m$ be the number of components of f that are not δ-coherent. Then f can be extended to a function $\tilde{f} : [0,1]^{n+p} \to [0,1]^m$ that is δ-coherent and such that $\forall i : 1 \leq i \leq m$:*

$$\forall x \in [0,1]^n, \forall c \in [0,1]^p, \qquad \tilde{f}_i(x,c) = f_i(x).$$

if f_i is δ-coherent on x.

Proof. Let $W = \{i_1, \ldots, i_p\} \subseteq \{1, \ldots, m\}$ be the set of indices of components of f that are not δ-coherent. For every $x \in [0,1]^n, c \in [0,1]^p$, we define $\tilde{f}_i(x,c) = f_i(x)$ if $i \notin W$, whereas if $i_j \in W$:

$$\tilde{f}_{i_j}(x,c) = \begin{cases} f_{i_j}(x) & \text{if } f_{i_j} \text{ is } \delta\text{-coherent on } x \\ c_j & \text{otherwise} \end{cases}$$

Then clearly $\tilde{f}$ extends f as specified in the statement. In addition, let $x \in [0,1]^n, c \in [0,1]^p$. By definition of $\tilde{f}$, for all $i \in \{1, \ldots, m\}$, either f_i is δ-coherent on x or is constant, and therefore $\delta \circ \tilde{f}_i(x,c) = \delta \circ f_i(x) = \delta \circ f_i \circ \delta(x) = \delta \circ \tilde{f}_i \circ \delta(x,c)$, or $\delta \circ \tilde{f}_i(x,c) = \delta \circ c_j(x) = \delta \circ c_j \circ \delta(x) = \delta \circ \tilde{f}_i \circ \delta(x,c)$ for any $j \in \{1, \ldots, p\}$. □

Basically, the idea behind the expansion $\tilde{f}$ is to "correct" the components that are not δ-coherent by adding one input for each of them.

5.2 Output Modification

The next theorem provides a criterion to associate a coherent function to a non-coherent one via output modification.

Theorem 4. *Let δ be a projection, $f, g\colon [0,1]^n \to [0,1]^m$, with g δ-coherent. The function $\hat{f}\colon [0,1]^n \to [0,1]^m$ defined by:*

$$\hat{f}_i(x) = \begin{cases} f_i(x) & \text{if } f_i \text{ is } \delta\text{-coherent on } x \\ g_i(x) & \text{otherwise} \end{cases}$$

for all $x \in [0,1]^n$ and $i = \{1, \dots, m\}$ is δ-coherent.

Proof. Let $x \in [0,1]^n$. Then for all $i \in \{1, \dots, m\}$, if f_i is coherent on x, then $\delta \circ \hat{f}_i(x) = \delta \circ f_i \circ \delta(x) = \delta \circ \hat{f}_i \circ \delta(x)$; otherwise $\delta \circ \hat{f}_i(x) = \delta \circ g_i(x) = \delta \circ g_i \circ \delta(x) = \delta \circ \hat{f}_i \circ \delta(x)$, since g_i is δ-coherent. □

A special case of this approach to make a function δ-coherent was already successfully used in (Barbiero et al. 2021). In this case, the explaining functions were constrained to be δ-coherent by defining g_i to be constantly equal to 1, demonstrating a gain of fidelity and accuracy of the model explainer forcing the δ-coherency with respect to an α-booleanization function.

5.3 General Case

The previous approaches can be seen as special instances of a general case by making explicit the mapping associating a δ-coherent to a fuzzy function.

Definition 4. Let δ be a projection. A *δ-coherency function* is a map Γ from the set $\mathcal{F}^h = \{f\colon [0,1]^n \to [0,1]^m \mid n, m \in \mathbb{N}\}$ to its subset of δ-coherent functions $\mathcal{C}^h_\delta$, such that $\Gamma(f) = f$ (idempotency) for every f that is δ-coherent.

Notice that requiring idempotency over δ-coherent functions is fundamental to define a well-suited δ-coherent function. Clearly, the cases of *domain extension* and *output modifications* lie under the scope of Definition 4; however, they are intrinsically not compositional, i.e., they are not well-behaved with respect to functional composition as captured by the following theorem.

Theorem 5. *Let δ be a projection, and let Γ be a δ-coherency function. Then Γ is not compositional, that is, there exists $f, g \in \mathcal{F}^h$ such that $\Gamma(g \circ f) \neq \Gamma(g) \circ \Gamma(f)$.*

Proof. Let $g\colon [0,1] \to [0,1]$ be a non-δ-coherent function. We have two cases:

Case domain of $\mathrm{Dom}(\Gamma(g)) \neq [0,1]$. The claim holds by taking $f\colon [0,1] \to [0,1]$ as the constant 0. In fact, $\Gamma(g \circ f)$ is defined while $\Gamma(g) \circ \Gamma(f)$ is not defined.

Case $\mathrm{Dom}(\Gamma(g)) = [0,1]$. Let $S \subset [0,1]$ be the subset of points in the domain of g, where g is not δ-coherent. Since g is not δ-coherent, $|S| > 0$. Furthermore, there exist $a, b \in S$, such that $\delta(a) = \delta(b)$ with $\delta(g(a)) \neq \delta(g(b))$ and $\delta(\Gamma(g)(a)) = \delta(\Gamma(g)(b))$, since $\Gamma(g)$ is δ-coherent by Definition 4. Hence, either $\Gamma(g)(a) \neq g(a)$ or $\Gamma(g)(b) \neq g(b)$. Let us assume, w.l.o.g., $\Gamma(g)(a) \neq g(a)$. Let $f\colon [0,1] \to [0,1]$ be the δ-coherent function constantly equal to a. Then clearly $g \circ f(x) = g(a)$, for all $x \in [0,1]$, and thus $g \circ f$ is δ-coherent (cf. Lemma 1). Then, by considering a, we have $\Gamma(g \circ f)(a) = (g \circ f)(a) = g(a) \neq \Gamma(g)(a) = (\Gamma(g) \circ f)(a) = (\Gamma(g) \circ \Gamma(f))(a)$. Hence Γ is not compositional.

□

5.4 From Fuzzy to δ-Coherent Functions

From Theorem 5 we see that witnessing of non-compositionality can be constructed for any given δ-coherency function Γ. This fact creates an obstacle to extending the explaining functor to non-δ-coherent functions. With this aim, we will define a new category that allows us to compositionally associate a δ-coherent function to any given fuzzy function. To do so, we have to first introduce an equivalence relation over the set of all fuzzy functions.

Definition 5. Let δ be a projection, Γ be a δ-coherency function, and $\mathcal{F}^h$ be the set of fuzzy functions. We denote by $\equiv_\Gamma$ the binary relation over $\mathcal{F}^h$ defined by:

$$f \equiv_\Gamma g \qquad \text{iff} \qquad \Gamma(f) = \Gamma(g).$$

Lemma 3. *Let δ be a projection, and let Γ be a δ-coherency function. Then:*

1. *$\equiv_\Gamma$ is an equivalence relation;*
2. *there exists a unique δ-coherent function f in every equivalence class of $\equiv_\Gamma$.*

Proof. First we can observe that (1) follows from the fact that the equality is an equivalence relation. For (2) suppose that f and g are δ-coherent functions in the same equivalence class, i.e., $\Gamma(f) = \Gamma(g)$. By Definition 4, $f = \Gamma(f) = \Gamma(g) = g$. □

With this tool, we can deal with the compositionality problem of non-δ-coherent functions. To this aim, we define the following category.

Theorem 6. *Let δ be a projection, and let Γ be a δ-coherency function. The pair $\mathcal{C}_{(\delta,\Gamma)} = (\mathcal{C}^o_{(\delta,\Gamma)}, \mathcal{C}^h_{(\delta,\Gamma)})$ where $\mathcal{C}^o_{(\delta,\Gamma)} = \{[0,1]^n \mid n \in \mathbb{N}\}$ and:*

1. *for all $A, B \in \mathcal{C}^o_{(\delta,\Gamma)}$, $Hom_{\mathcal{C}_{(\delta,\Gamma)}}(A, B) = \{[f]_{\equiv_\Gamma} \mid \Gamma(f)\colon A \to B \mid f \in \mathcal{F}^h\}$, where $[f]_{\equiv_\Gamma}$ is the $\equiv_\Gamma$-equivalence class of f;*
2. *for all $A, B, C \in \mathcal{C}^o_{(\delta,\Gamma)}$, $[f]_{\equiv_\Gamma} \in Hom_{\mathcal{C}_{(\delta,\Gamma)}}(A, B)$, and $[g]_{\equiv_\Gamma} \in Hom_{\mathcal{C}_{(\delta,\Gamma)}}(B, C)$, we define $[g]_{\equiv_\Gamma} \circ [f]_{\equiv_\Gamma} = [\Gamma(g) \circ \Gamma(f)]_{\equiv_\Gamma}$;*

forms a category that we call the category of quotient fuzzy functions.

Proof. First we observe that $\circ$ is well-defined since, by Definition 4 and Lemma 3, for all $[f]_{\equiv_\Gamma} \in \mathrm{Hom}_{\mathcal{C}_{(\delta,\Gamma)}}(A, B)$, $[g]_{\equiv_\Gamma} \in \mathrm{Hom}_{\mathcal{C}_{(\delta,\Gamma)}}(B, C)$, and for all $f' \in [f]_{\equiv_\Gamma}$ and $g' \in [g]_{\equiv_\Gamma}$, $\Gamma(f)$ and $\Gamma(g)$ are the unique δ-coherent functions in $[f]_{\equiv_\Gamma}$ and $[g]_{\equiv_\Gamma}$ respectively, and thus $[g]_{\equiv_\Gamma} \circ [f]_{\equiv_\Gamma} = [\Gamma(g) \circ \Gamma(f)]_{\equiv_\Gamma} = [\Gamma(g') \circ \Gamma(f')]_{\equiv_\Gamma} = [g']_{\equiv_\Gamma} \circ [f']_{\equiv_\Gamma}$. Let us consider $X = [0,1]^n \in \mathcal{C}^o_\delta$, then $[\mathrm{Id}_X]_{\equiv_\Gamma}$, the $\equiv_\Gamma$-class of the identity over X, is the identity morphism since Id_X is the δ-coherent function in $[\mathrm{Id}_X]_{\equiv_\Gamma}$. Thus $[\mathrm{Id}_X]_{\equiv_\Gamma} \in \mathrm{Hom}_{\mathcal{C}_{(\delta,\Gamma)}}(X, X)$.

Let us assume $[f]_{\equiv_\Gamma} \in \mathrm{Hom}_{\mathcal{C}_{(\delta,\Gamma)}}([0,1]^n,[0,1]^m)$ and $[g]_{\equiv_\Gamma} \in \mathrm{Hom}_{\mathcal{C}_{(\delta,\Gamma)}}([0,1]^m,[0,1]^k)$. Then we have to prove that $[g]_{\equiv_\Gamma} \circ [f]_{\equiv_\Gamma} \in \mathrm{Hom}_{\mathcal{C}_{(\delta,\Gamma)}}([0,1]^n,[0,1]^k)$. For this purpose, it is enough to show that $\Gamma(g) \circ \Gamma(f)$ is δ-coherent, but this was already observed in the proof of Theorem 1. Thus $[g]_{\equiv_\Gamma} \circ [f]_{\equiv_\Gamma} = [\Gamma(g) \circ \Gamma(f)]_{\equiv_\Gamma} \in \mathrm{Hom}_{\mathcal{C}_{(\delta,\Gamma)}}([0,1]^n,[0,1]^k)$. □

We can now define a functor from $\mathcal{C}_{(\delta,\Gamma)}$ to $\mathcal{C}_\delta$.

Definition 6. Let $\delta\colon [0,1] \to \{0,1\}$ be a projection, Γ a δ-coherency function, and $F_\Gamma : \mathcal{C}_{(\delta,\Gamma)} \to \mathcal{C}_\delta$ the mapping defined by:

1. for every $X = [0,1]^n \in \mathcal{C}^o_{(\delta,\Gamma)}$, $F_\Gamma(X) = [0,1]^n \in \mathcal{C}^o_\delta$, i.e., $F_\Gamma(X) = X$;
2. for all $[f]_{\equiv_\Gamma} \in \mathrm{Hom}_{\mathcal{C}_{(\delta,\Gamma)}}([0,1]^n,[0,1]^k)$ we define $F_\Gamma([f]_{\equiv_\Gamma}) = \Gamma(f) \in \mathcal{C}^h_\delta$.

Theorem 7. *F_Γ introduced in Definition 6 is a functor from $\mathcal{C}_{(\delta,\Gamma)}$ to $\mathcal{C}_\delta$.*

Proof. First, we have to prove that for every object $A \in \mathcal{C}^o_{(\delta,\Gamma)}$, $F_\Gamma([\mathrm{Id}_A]_{\equiv_\Gamma}) = \mathrm{Id}_{F_\Gamma(A)} = \mathrm{Id}_A$, which is trivial as $F_\Gamma([\mathrm{Id}_A]_{\equiv_\Gamma}) = \Gamma(\mathrm{Id}_A) = \mathrm{Id}_A$, as Id_A is δ-coherent. Secondly, let $[f]_{\equiv_\Gamma} \in \mathrm{Hom}_{\mathcal{C}_{(\delta,\Gamma)}}([0,1]^n,[0,1]^m)$ and $[g]_{\equiv_\Gamma} \in \mathrm{Hom}_{\mathcal{C}_{(\delta,\Gamma)}}([0,1]^m,[0,1]^k)$ we have to prove that $F_\Gamma([g]_{\equiv_\Gamma} \circ [f]_{\equiv_\Gamma}) = F_\Gamma([g]_{\equiv_\Gamma}) \circ F_\Gamma([f]_{\equiv_\Gamma})$. However, this follows from Theorem 6. □

Notice that, as the composition of two functors is a functor, trivially by composing the functor F_Γ and F_δ, we get a functor from $\mathcal{C}_{(\delta,\Gamma)}$ to $\mathcal{B}$. This functor, which we indicate by $F_{(\delta,\Gamma)}$, is characterized as follows.

Theorem 8. *Let $\delta\colon [0,1] \to \{0,1\}$ be a projection and Γ a δ-coherency function. Then the mapping $F_{(\delta,\Gamma)} : \mathcal{C}_{(\delta,\Gamma)} \to \mathcal{B}$ defined by:*

1. *for every $X = [0,1]^n \in \mathcal{C}^o_{(\delta,\Gamma)}$, $F_{(\delta,\Gamma)}(X) = \{0,1\}^n \in \mathcal{B}^o$;*
2. *for every morphism $[f]_{\equiv_\Gamma} \in Hom_{\mathcal{C}_{(\delta,\Gamma)}}([0,1]^n,[0,1]^m)$ the function $F_{(\delta,\Gamma)}(f) = \Gamma(f)^\delta \in \mathcal{B}^h$, with $\Gamma(f)^\delta\colon \{0,1\}^n \to \{0,1\}^m$ such that $\Gamma(f)^\delta = \delta \circ \Gamma(f)$*

is a functor from $\mathcal{C}_{(\delta,\Gamma)}$ to $\mathcal{B}$, and it satisfies $F_{(\delta,\Gamma)} = F_\delta \circ F_\Gamma$.

Proof. The proof follows from Theorem 2 and Theorem 7. □

The results presented in this section can be fully generalized to arbitrary sets beyond $[0,1]$ and $\{0,1\}$, thereby allowing the use of different types of logic as targets for a coherent explainer. Furthermore, the functions Γ and δ are subject to no additional constraints apart from the idempotency conditions on δ-coherent functions and the image of δ. Specifically, $\Gamma(f) = f$ for any δ-coherent f and $\delta(x) = x$ for all $x \in \delta([0,1])$. This demonstrates that our approach possesses significant flexibility and can be uniformly applied to explainers operating on non-Boolean logics.

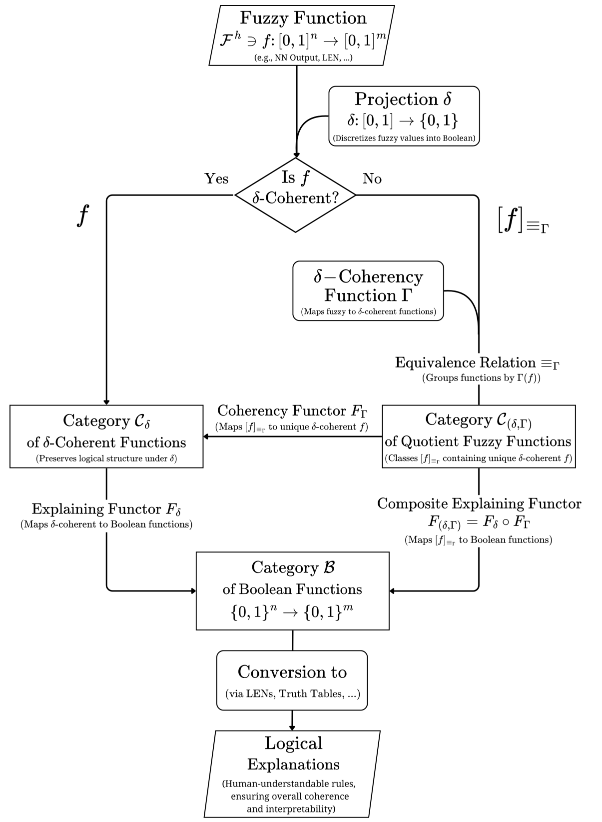

Figure 1: Illustration of our proposed categorical framework for generating coherent logical explanations from fuzzy functions.

Explaining fuzzy functions by categorial functors. As we noticed, it is not possible to define a functor from $\mathcal{F}$ to $\mathcal{B}$ that preserves coherency. However, we can exploit the explaining functors defined above to associate to each fuzzy function a Boolean function and then a Boolean formula that can be taken as an explanation of it. Given a fuzzy function $f \in \mathcal{F}^h$, with $f : [0,1]^n \to [0,1]^m$, a projection δ, and a δ-coherency function Γ, this can be done according to the following steps:

1. We exploit the map Γ that associates to f a unique δ-coherent function $\Gamma(f)$ in its equivalence class $[f]_{\equiv_\Gamma}$.
2. Then, by means of the functor F_δ, we obtain the Boolean function $\Gamma(f)^\delta \in \mathcal{B}^h$.
3. Finally, each component of $\Gamma(f)^\delta$ can be written as a Boolean formula by converting its truth-table into a disjunctive logic clause.

The overall categorical framework for deriving coherent logical explanations is presented in Figure 1.

6 Experimental Analysis

To demonstrate the practical validity of our proposed approach, we present experiments on synthetic benchmarks assessing (1) learning a naturally δ-coherent function (Boolean XOR), and (2) learning a inherently non-δ-coherent fuzzy function (Łukasiewicz t-conorm). We assess

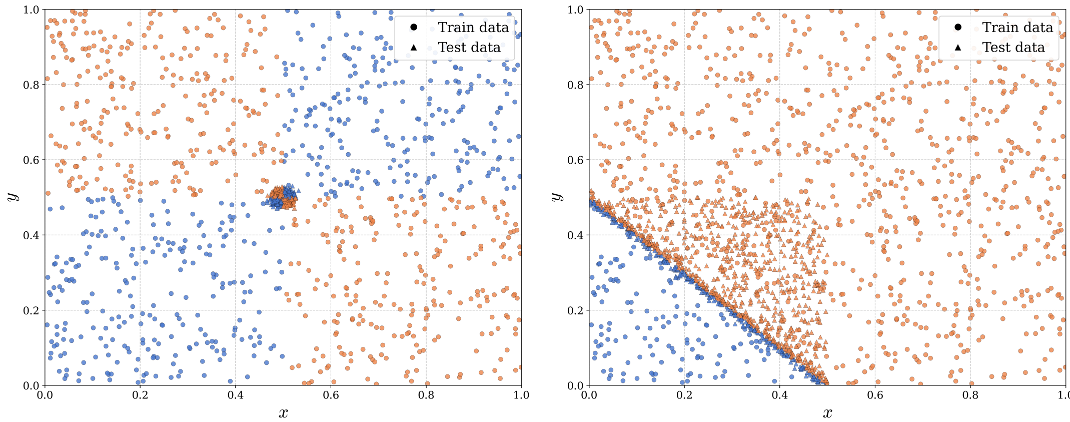

Figure 2: Data distributions for the XOR (left) and fuzzy OR (right) experiments, with colors indicating class labels.

Exp.		Accuracy			Coherency			Explanations		
		Train	Val.	Test	Train	Val.	Test	l	FOL	Accuracy
(1)	$f^{(1)}$	100	100	95.5	100	100	94.8	0	$(x \wedge y) \vee (\neg x \wedge \neg y)$	94.8
								1	$(x \wedge \neg y) \vee (y \wedge \neg x)$	94.8
(2)	$f^{(2)}$	99.9	99.2	88.4	98.5	97.6	67.1	0	$\neg y$	67.1
								1	$x \vee \neg x = 1$	75.7
	$\hat{f}^{(2)}$							0	$\neg x \wedge \neg nc$	83.8
								1	$x \vee nc$	83.8

Table 1: Experimental results and explanations on the test set for (1) (XOR) and (2) (fuzzy OR) experimental settings. Values are in percentages over all corresponding data points. Explanations are computed on the test set.

the resulting coherence, provide logic-based explanations, and, for the non-coherent case, we extend the explaining functor producing coherent explanations.

Experimental settings. For the experiments, we consider the following target functions:

(1) $f^{(1)}: [0,1]^2 \to \{0,1\}$ as the boolean XOR function $f^{(1)}(x,y) = \delta(x) \oplus \delta(y)$;

(2) $f^{(2)}: [0,1]^2 \to [0,1]$ as the Łukasiewicz t-conorm (fuzzy OR) $f^{(2)}(x,y) = \min(1, x+y)$.

The function $f^{(1)}$ is inherently $\delta_{0.5}$-coherent, while $f^{(2)}$ is non-δ-coherent in the region $W = \{(x,y) \mid x+y \geq 0.5, x \leq 0.5, y \leq 0.5\} \subset [0,1]^2$.

Both functions are implemented with a LEN trained on synthetic datasets of uniformly generated points $(x,y) \in [0,1]^2$, comprising 1000 training and 250 validation samples for binary classification ($l \in \{0,1\}$). Performance evaluation uses a test set of 1000 points strategically concentrated in (1) around the central decision boundary or (2) around the region W (see Figure 2).

We used PReLU activation functions between layers, with hyperparameters (number and size of hidden layers, learning rate, weight decay, coherence regularization multiplier) optimized through random search for the best validation accuracy with early stopping.

Results. Table 1 reports the results grouped for both experiments, including accuracy, coherence between concepts and labels, FOL explanations, and their quality.

LENs predict output classes while providing FOL explanations in terms of their inputs. In experiment (1), the inputs are concepts computed by an embedding layer, and the coherence between the Boolean concepts and the labels was encouraged via regularization. For (2), we extended the explaining functor by applying Theorem 3, yielding a post-hoc modified explainer $\hat{f}^{(2)}$ that discerns non-coherent samples by adding an additional feature $nc \in \{0,1\}$.

The experiments reveal that when the underlying function is naturally δ-coherent, as in (1), a LEN achieves high accuracy and generates logically consistent, high-quality FOL explanations mirroring the target function. Conversely, for the non-coherent fuzzy OR function in (2), while model accuracy remains high, the direct FOL explanation ('$x \vee \neg x = 1$' for class 1) is trivial and accuracy drops significantly. Applying the coherency function ($\hat{f}^{(2)}$) to generate explanations recovers accuracy (83.8%) by explicitly accounting for non-coherence, demonstrating the practical utility of ensuring δ-coherence for faithful explanations and aligning with the principles motivating our explaining functors.

7 Conclusions and Future Work

This paper presents a novel categorical framework for producing logic-based post-hoc explanations. Using category theory, our approach offers a mathematically rigorous method for modeling different learning architectures, ensuring both fidelity and compositional coherence. In particular, we introduce a new theory for generating explanations founded on the concept of explaining functors and supported by mathematical proofs, that enforces structural consistency even in complex pipelines.

Our findings point to several promising avenues for future research. A primary direction involves the empirical validation of coherent Boolean explanations within concept-based models, such as Concept Bottleneck Models. More broadly, the proposed framework is applicable to any model that aggregates input features into discrete, interpretable representations in order to generate logical or symbolic explanations—particularly when such models are composed in sequence. This applicability extends to hierarchical and causal variants of CBMs. An additional contribution of this work is the establishment of a principled foundation that integrates theoretical rigor with practical interpretability, thereby advancing the state of explainable AI. In this context, we anticipate that analogous categorical constructions—such as specifically designed functors—could be developed to rigorously investigate alternative explanation techniques, including those grounded in perturbation analysis or feature-importance measures.

Acknowledgments

We recognize the support of the following projects and research grants. S. Fioravanti, P. Frazzetto, R. Confalonieri and N. Navarin acknowledge the support of “Future AI Research (FAIR) - Spoke 2 Integrative AI - Symbolic conditioning of Graph Generative Models (SymboliG)” funded by the European Union under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of March 15, 2022 of Italian Ministry of University and Research – NextGenerationEU, Code PE0000013, Concession Decree No. 1555 of October 11, 2022 CUP C63C22000770006, S. F. also acknowledges the Charles University under the grant number PRIMUS/24/SCI/008. R. Confalonieri also acknowledges funding support from the “NeuroXAI” project (BIRD231830). F. Zanasi acknowledges support from ARIA’s Safeguarded AI Programme. This work has also been partially supported by the Partnership Extended PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI” and ERC-2018-ADG G.A. 834756 “XAI: Science and technology for the eXplanation of AI decision making”. P. Barbiero acknowledges support from the Swiss National Science Foundation project “IMAGINE” (No. 224226).

Contribution Statement

The first two authors contributed equally to this work.

References

Abramsky, S., and Coecke, B. 2004. A categorical semantics of quantum protocols. In *19th Annual IEEE Symposium on Logic in Computer Science*. IEEE.

Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Ser, J. D.; Díaz-Rodríguez, N.; and Herrera, F. 2023. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion* 99(C):101805.

Arenas, M.; Baez, D.; Barceló, P.; Pérez, J.; and Subercaseaux, B. 2021. Foundations of symbolic languages for model interpretability. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Audemard, G.; Bellart, S.; Bounia, L.; Koriche, F.; Lagniez, J.-M.; and Marquis, P. 2021. On the Computational Intelligibility of Boolean Classifiers. In *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning*, 74–86.

Barbiero, P.; Ciravegna, G.; Georgiev, D.; and Giannini, F. 2021. Pytorch, explain! a python library for logic explained networks. *arXiv preprint arXiv:2105.11697*.

Barbiero, P.; Ciravegna, G.; Giannini, F.; Lió, P.; Gori, M.; and Melacci, S. 2022. Entropy-based logic explanations of neural networks. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, 6046–6054. AAAI Press.

Bobillo, F.; Bobed, C.; Mena, E.; and Straccia, U. 2024. A fuzzy logic-based approach to semantic query answering with missing values. In *2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8. IEEE.

Ciravegna, G.; Giannini, F.; Gori, M.; Maggini, M.; and Melacci, S. 2020. Human-driven fol explanations of deep learning. In *Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}*, 2234–2240. International Joint Conferences on Artificial Intelligence Organization.

Ciravegna, G.; Barbiero, P.; Giannini, F.; Gori, M.; Lió, P.; Maggini, M.; and Melacci, S. 2023. Logic explained networks. *Artificial Intelligence* 314:103822.

Colombini, J. J.; Bonchi, F.; Giannini, F.; Giannotti, F.; Pellungrini, R.; Frosini, P.; et al. 2025. Mathematical foundation of interpretable equivariant surrogate models. In *The 3nd World Conference on eXplainable Artificial Intelligence, XAI-2025*.

Cruttwell, G. S.; Gavranović, B.; Ghani, N.; Wilson, P.; and Zanasi, F. 2022. Categorical foundations of gradient-based learning. In *European Symposium on Programming*.

Davies, A.; Veličković, P.; Buesing, L.; Blackwell, S.; Zheng, D.; Tomašev, N.; Tanburn, R.; Battaglia, P.; Blundell, C.; Juhász, A.; et al. 2021. Advancing mathematics by guiding human intuition with ai. *Nature* 600(7887):70–74.

Debot, D.; Barbiero, P.; Dominici, G.; and Marra, G. 2025. Interpretable hierarchical concept reasoning through attention-guided graph learning. *arXiv preprint arXiv:2506.21102*.

Dominici, G.; Barbiero, P.; Giannini, F.; Gjoreski, M.; Langeinrich, M.; et al. 2024. Anycbms: How to turn any black box into a concept bottleneck model. In *CEUR WORKSHOP PROCEEDINGS*, volume 3793, 81–88. CEUR-WS.

Dominici, G.; Barbiero, P.; Zarlenga, M. E.; Termine, A.; Gjoreski, M.; Marra, G.; and Langheinrich, M. 2025. Causal concept graph models: Beyond causal opacity in deep learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Duneau, T.; Bruhn, S.; Matos, G.; Laakkonen, T.; Saiti, K.; Pearson, A.; Meichanetzidis, K.; and Coecke, B. 2024. Scalable and interpretable quantum natural language processing: an implementation on trapped ions. *arXiv preprint arXiv:2409.08777*.

Ghorbani, A.; Wexler, J.; Zou, J. Y.; and Kim, B. 2019. Towards automatic concept-based explanations. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 9273–9282.

Giannini, F.; Fioravanti, S.; Keskin, O.; Lupidi, A. M.; Magister, L. C.; Lió, P.; and Barbiero, P. 2023. Interpretable graph networks formulate universal algebra conjectures. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Giannini, F.; Fioravanti, S.; Barbiero, P.; Tonda, A.; Liò, P.; and Lavore, E. D. 2024. Categorical foundation of explainable AI: A unifying theory. In Longo, L.; Lapuschkin, S.; and Seifert, C., eds., *Explainable Artificial Intelligence - Second World Conference, xAI 2024, Valletta, Malta, July 17-19, 2024, Proceedings, Part III*, volume 2155 of *Communications in Computer and Information Science*, 185–206. Springer.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Naretto, F.; Turini, F.; Pedreschi, D.; and Giannotti, F. 2024. Stable and actionable explanations of black-box models through factual and counterfactual rules. *Data Min. Knowl. Discov.* 38(5):2825–2862.

Hájek, P. 2013. *Metamathematics of fuzzy logic*, volume 4. Springer Science & Business Media.

Jain, R.; Ciravegna, G.; Barbiero, P.; Giannini, F.; Buffelli, D.; and Lió, P. 2022. Extending logic explained networks to text classification. In *Empirical Methods in Natural Language Processing*. [Under review].

Jiménez-Luna, J.; Grisoni, F.; and Schneider, G. 2020. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* 2(10):573–584.

Keskin, O.; Lupidi, A.; Giannini, F.; Fioravanti, S.; Magister, L. C.; Barbiero, P.; and Liò, P. 2023. Bridging equational properties and patterns on graphs: an ai-based approach. In *Proceedings of Machine Learning Research*, volume 221, 156 – 168.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 5338–5348. PMLR.

Letham, B.; Rudin, C.; McCormick, T. H.; and Madigan, D. 2015. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9(3):1350 – 1371.

Mac Lane, S. 1978. *Categories for the Working Mathematician*. Graduate Texts in Mathematics. Springer New York.

Marques-Silva, J., and Ignatiev, A. 2022. Delivering trustworthy ai through formal xai. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(11):12342–12350.

Marques-Silva, J. 2024. Logic-based explainability: Past, present and future. In *Leveraging Applications of Formal Methods, Verification and Validation. Software Engineering Methodologies: 12th International Symposium, ISoLA 2024, Crete, Greece, October 27–31, 2024, Proceedings, Part IV*, 181–204. Berlin, Heidelberg: Springer-Verlag.

Palacio, S.; Lucieri, A.; Munir, M.; Ahmed, S.; Hees, J.; and Dengel, A. 2021. Xai handbook: Towards a unified framework for explainable ai. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3766–3775.

Poeta, E.; Ciravegna, G.; Pastor, E.; Cerquitelli, T.; and Baralis, E. 2023. Concept-based explainable artificial intelligence: A survey. *arXiv preprint arXiv:2312.12936*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In McIlraith, S. A., and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 1527–1535. AAAI Press.

Rodatz, B.; Fan, I.; Laakkonen, T.; Ortega, N. J.; Hoffman, T.; and Wang-Mascianica, V. 2024. A pattern language for machine learning tasks. *CoRR* abs/2407.02424.

Schmidt, M., and Lipson, H. 2009. Distilling free-form natural laws from experimental data. *science* 324(5923):81–85.

Shiebler, D.; Gavranović, B.; and Wilson, P. 2021. Category theory in machine learning. *arXiv preprint arXiv:2106.07032*.

Stein, D., and Staton, S. 2021. Compositional semantics for probabilistic programs with exact conditioning. In *36th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2021, Rome, Italy, June 29 - July 2, 2021*, 1–13. IEEE.

Tull, S.; Lorenz, R.; Clark, S.; Khan, I.; and Coecke, B. 2024. Towards compositional interpretability for XAI. *CoRR* (abs/2406.17583).