Explaining Black-box Classifiers: Properties and Functions

Leila Amgoud
CNRS – IRIT, France
amgoud@irit.fr

Abstract

Explaining black-box classification models is a hot topic in AI, with the overall goal of improving trust in decisions made by such models. Several works have been done and diverse functions have been proposed. However, their formal properties and links have not been sufficiently studied. This paper presents four contributions: The first consists of investigating global explanations of black-box classifiers. We provide a formal and unifying framework in which such explanations are defined from the whole feature space. The framework is based on two concepts, which are seen as two types of global explanations: arguments in favour of (or pro) predictions and arguments against (or con) predictions. The second contribution consists of defining various types of local explanations (abductive explanations, counterfactuals, contrastive explanations) from the whole feature space, investigating their properties, links and differences, and showing how they relate to global explanations. The third contribution consists of analysing and defining explanation functions that generate (global, local) abductive explanations from incomplete information (i.e., from a subset of the feature space). We start by proposing two desirable properties that an explainer would satisfy, namely *success* and *coherence*. The former ensures the existence of explanations while the latter ensures their correctness. We show that in the incomplete case, the two properties cannot be satisfied together. The fourth contribution consists of proposing two functions that generate abductive explanations and which satisfy coherence at the expense of success.

Keywords: Classification, Explainability, Arguments.

1. Introduction

In recent years, remarkable advances have been made in data-driven artificial intelligence in general, and deep learning in particular. In this sub-field of AI, the idea is to learn a targeted objective (eg. the class of an object) from a vast quantity of data. Despite the recent successes, existing models are opaque, in that their predictions (outcomes) can hardly be explained in a transparent way. Indeed, the predictions of these models resist analysis due to their inherent non-linear behaviour and their vast amount of interacting parameters. This opacity is seen as a great limitation, which impedes the relevance of those models from a theoretical point of view, since their properties are difficult to investigate, and from a practical point of view, as many applications, such as healthcare or embedded systems need guarantees to be deployed, and others, e.g in the legal or financial domain require transparency to be accepted. Hence, improving trust in decisions made by such models becomes crucial for the acceptance of automated systems, and an important way of doing that is by providing explanations for the outcomes of the models. Explanations help human users understand i) why a decision was reached and why alternative decisions were not recommended, ii) what could be changed to receive a desired outcome in the future. They may also help designers or data scientists to detect possible flaws of models. Interested readers can refer to [1, 2, 3, 4, 5, 6] for more information on explainability.

Explaining the functionality of classification models and their rationale has become a vital need, and has generated a lot of research (see [7, 2, 1] for recent surveys on explainability of machine learning models). Existing *explanation functions*, or *explainers*, can be classified in three different ways. The first way distinguishes explainers that provide explanations for individual predictions (i.e., explaining the decision of a given instance), called *local explanations* (eg. [8, 9, 10, 11, 12]) from explainers that provide explanations for classes independently of instances, called *global explanations*, (eg. [13, 14]). The second way is based on the information used for generating explanations. Some explainers, like those studied in [11, 13, 14, 15, 16], use the whole set of instances, called the *feature space*, while others like Anchors [9], LIME [8] and the non-monotonic explainer from [17], use only subsets of the feature space. The third way distinguishes explainers which look inside the model from those which consider a model as a black-box whose internal reasoning is left unspecified. The former provide insight into the internal decision-making process, e.g., [16, 18, 19, 20]. They are suitable for explaining interpretable models like decision trees and Bayesian networks; however they may not be feasible for complex and non-interpretable ones like deep neural networks. Furthermore, a predominant finding from research in philosophy and social science is that a full causal explanation of a prediction is not desirable for humans, as they do not need to understand the algorithm followed by a model. In other words, an explanation does not necessarily hinge on the general public understanding of how algorithmic systems function. Hence, the second family of explainers considers a classification model as a black-box and provides explanations without looking inside the model. It looks for possible correlations between input data and the predictions made by a model. This approach has been applied to non-interpretable models (eg. [9, 10, 14, 21, 22, 23, 24]) and also to interpretable ones (eg. [11, 12, 19]).

Throughout the paper, we follow this second line of research and thus consider black-box classification models. In this context, existing literature suffers from three main shortcomings. The first concerns its focus mainly on local explanations, indeed existing works provide explanations for individual instances. Such explanations are important in particular for providing feedback for the users of a classifier (eg. explaining why a loan has been rejected for a given customer). However, the compatibility of these local explanations with the global behaviour of a model has not been sufficiently studied. More generally, the question of explaining classes (instead of instances) received little interest. Such explanations may be important for data scientists to understand how a model assigns classes to input data.

It is also worth mentioning that several types of explanations have been defined. The most prominent ones are *prime implicants* [18], called also *abductive explanations* in [14], *counterfactuals* [24], *semi-factual* [25] and *contrastive* explanations [22, 10]. The former are seen as sufficient reasons for getting an outcome, counterfactuals are changes that should occur for avoiding an outcome, and contrastive explanations clarify why an outcome is proposed instead of another (desirable) one [6]. The second shortcoming of this literature is that, with a rare exception [11], the previous types of explanations have been studied in restrictive, practical, experimental cases, and mostly in isolation. A comprehensive formal comparison is thus missing while it is crucial for clarifying the role and added value of each notion, and relate them to desirable properties for explanations.

The third shortcoming is due to the information used for generating explanations. Some works like [11, 12, 16] generate abductive explanations by exploring the whole feature space (i.e., complete information), which may not be reasonable in practice especially for complex models. Other explainers, like Anchors [9] and LIME [8], generate explanations from a proper subset of the feature space (i.e., incomplete information). The two approaches (complete/incomplete information) have not been compared neither formally nor in an experimental way, and the strengths and weaknesses of each approach are still unclear.

This paper bridges the above three gaps with four-fold contributions. The first contribution consists of investigating global explanations of black-box classifiers. We provide a formal and unifying framework in which such explanations are defined from the whole feature space. The framework is based on two concepts, which are seen as two types of global explanations: arguments in favour of (or pro) predictions and arguments against (or con) predictions. The former justify why a class is suggested by a classifier and the latter state why a class is not proposed. We investigate the formal properties of both types of arguments and show that they are dual. Indeed, we provide ways for generating arguments pro a class from arguments con the class and vice versa.

The second contribution consists of defining various types of local explanations (abductive explanations, counterfactuals, contrastive explanations) from the whole feature space, investigating their properties, links and differences, and showing how they relate to global explanations. We show that abductive explanations are based on arguments pro while contrastive explanations and counterfactuals are based on arguments con.

The third contribution consists of analysing and defining explanation functions that generate (global, local) abductive explanations from incomplete information, i.e. from a proper subset of the feature space. The subset can be chosen in different ways: It may be the neighbourhood of specific instances, or a set of instances on which the classifier returns a good confidence, or simply a dataset on which it was trained. We start by proposing two desirable properties that an explainer would satisfy, namely *success* and *coherence*. The former ensures the existence of explanations while the latter ensures their correctness. Indeed, it forbids two compatible sets of features-values from explaining distinct classes. Then, we introduce *plausible* explanations, which are abductive explanations generated from a subset of instances. Such explanations are only plausible since they are generated from incomplete information, hence they may no longer be valid if the subset of instances is extended with further ones. Such explainers are thus non-monotonic. Furthermore, they violate the property of coherence leading to incorrect explanations, elucidating thus the origin of the flaws of the functions Anchors and LIME. We also show that any explainer which selects a subset of plausible explanations violates either coherence or success. In other words, an explainer which generates a proper subset of plausible explanations cannot satisfy the two properties (success and coherence) together. Defining non-monotonic explainers that are coherent remains a challenge in the XAI literature. The fourth contribution bridges this gap by proposing two functions that generate abductive explanations and which satisfy the coherence property. They are based on argumentation, a reasoning approach which is based on the construction and evaluation of interacting arguments (see [26] for more on argumentation theory and its applications).

Argumentation is based on the justification of claims by *arguments*, i.e. reasons for accepting claims. It received great interest from the Artificial Intelligence community since late 1980s, namely as a unifying approach for nonmonotonic reasoning [27]. It was later used for solving different other problems like reasoning with inconsistent information (eg. [28, 29]), decision making (eg. [30**?**]), classification (eg. [31]), etc. It has also several practical applications, namely in legal and medical domains (see [32]). Whatever the problem to be solved, an argumentation process follows generally four main steps: to justify claims by arguments, identify (attack, support) relations between arguments, evaluate the arguments, and define an output. The last step depends on the results of the evaluation. For instance, an inference system draws formulas that are justified by what is qualified at the evaluation step as "strong" arguments. Evaluation of arguments is thus crucial as it impacts the outcomes of argument-based systems. Consequently, a plethora of methods, called *semantics*, have been proposed in the literature. The very first ones are the *extension* semantics (*stable, preferred, complete* and *grounded*) that have been proposed by Dung in [33]. Argumentation is a powerful approach for solving different kinds of conflicts. In this paper, conflicts are due to the violation of the coherence property. We define then two argumentation systems which solve the conflicts, each of which defines an explanation function that guarantees coherence at the expense of success.

This paper unifies and extensively develops the content of two conference papers [13, 17]. It contains detailed proofs of all the results and investigates more deeply than [13] the notions of local and global explanations under complete information. It also extends [17] in different ways. It introduces the novel property of success and shows an impossibility result stating that there is no function that generates abductive explanations from a subset of the feature space and satisfies both coherence and success. The paper studies deeply the properties of the explainer proposed in [17] and introduces a novel one which takes into account priorities among features.

The paper is structured as follows: Section 2 presents a background on classification and some useful notations. Section 3 defines two types of explanation functions and some properties. Section 4 investigates global and local explanations under complete information while Section 5 studies the case of incomplete information. Section 6 is devoted to related work and the last section to concluding remarks and perspectives. Proofs are put in an appendix at the end of the document.

2. Classification Problem

Let $\mathcal{F} = \{f_1, \ldots, f_n\}$ be a finite and non-empty set of *features* (called also *attributes*), $\texttt{dom}$ is a function on $\mathcal{F}$ such that, for every $f \in \mathcal{F}$, $\texttt{dom}(f)$ is countable (discrete domains) with $|\texttt{dom}(f)| > 1$. Throughout the paper, we focus on the important case of discrete features, as is the case in important applications (e.g., image or natural language processing). It is worth mentioning that there are well-known ways of discretizing continuous attributes, which sometimes gives better results on some learning algorithms, see [34]. We call a *literal* any pair (f, v) where $f \in \mathcal{F}$ and $v \in \texttt{dom}(f)$, and we denote by $\texttt{Lit}$ the set of all possible literals. A set of literals is *consistent* if it does not contain two literals that assign distinct values to the same feature.

Definition 1 (Consistency). *A set $H \subseteq \texttt{Lit}$ is* consistent *iff $\nexists (f, v)$, $(f', v') \in H$ such that $f = f'$ and $v \neq v'$. Otherwise, H is said to be* inconsistent*.*

We call an *instance* any subset of literals in which every feature (of the set $\mathcal{F}$) appears exactly once, i.e., it is an assignment of values to all features. We denote by $\texttt{Inst}$ the set of all instances and call it *feature space*. Clearly, the set $\texttt{Inst}$ is finite since $\mathcal{F}$ and $\texttt{dom}$ are finite. Furthermore, its elements are all consistent and any consistent set of literals is included in at least one instance.

Property 1. *The following properties hold.*

1. *For any $x \in \texttt{Inst}$, x is consistent.*
2. *For any $H \subseteq \texttt{Lit}$ such that H is consistent, the following hold:*
   (a) *$\forall H' \subset H$, H' is consistent.*
   (b) *$\exists x \in \texttt{Inst}$ such that $H \subseteq x$.*

For $x \in \texttt{Inst}$, $H \subseteq \texttt{Lit}$ such that H is consistent, $x_{\downarrow H}$ denotes the set of literals obtained by replacing in x the values of the common features to the two sets by those in H and keeping the values of the remaining (non-common) features unchanged. Formally:

$$x_{\downarrow H} = H \ \cup \ \{(f, v) \mid (f, v) \in x \text{ and } \nexists (f, v') \in H\}.$$

It is easy to show that the modified version of x is itself an instance, i.e., element of `Inst`.

Property 2. *For any $x \in \texttt{Inst}$, for any $H \subseteq \texttt{Lit}$, if H is consistent, then $x_{\downarrow H} \in \texttt{Inst}$.*

Let $\mathcal{C} = \{c_1, \ldots, c_m\}$, with $m > 1$, be a finite and non-empty set of possible distinct *classes*.

Definition 2 (Theory). *A classification* theory *is a tuple $\texttt{T} = \langle \mathcal{F}, \texttt{dom}, \mathcal{C} \rangle$.*

A *classification model* or *classifier* is a function which assigns to every instance $x \in \texttt{Inst}$ of a theory $\texttt{T} = \langle \mathcal{F}, \texttt{dom}, \mathcal{C} \rangle$ a single prediction, which is a class from the set $\mathcal{C}$. In some classification tasks, an instance can be assigned several classes, however this case is beyond the scope of this paper.

Definition 3 (Classification Model). *Let $\texttt{T} = \langle \mathcal{F}, \texttt{dom}, \mathcal{C} \rangle$ be a theory. A* classification model *is a function $\texttt{F} : \texttt{Inst} \rightarrow \mathcal{C}$.*

Let us illustrate the above notions with a simple example borrowed from [11].

Example 1. *Consider the task of college admission. Decisions are made on the basis of four binary features: Entrance exam (E), First time entrance (F), Work experience (W), GPA (G). The decision is binary: a candidate is either admitted or denied, so $\mathcal{C} = \{\texttt{Admitted}, \texttt{Denied}\}$. Consider a Bayesian network classifier `F` whose reasoning is represented by the following rules:*

- *If $E = 1$ and $F = 0$, then* `Admitted`*,*
- *If $E = 1$, $F = 1$, $W = 1$, then* `Admitted`*,*
- *If $E = 1$, $F = 1$ and $W = 0$ and $G = 1$, then* `Admitted`*,*
- *If $E = 1$, $F = 1$ and $W = 0$ and $G = 0$, then* `Denied`*,*
- *If $E = 0$, then* `Denied`*.*

Assume a student who passed the entrance exam, is a first-time applicant, has no work experience and a high GPA. This corresponds to the instance $x = \{(E, 1), (F, 1), (W, 0), (G, 1)\}$ *with* $\mathtt{F}(x) = \mathtt{Admitted}$*.*

The previous example uses an interpretable classifier, Bayesian network. The following example uses a complex classifier, like a deep neural network.

Example 2. *Assume a classification problem of deciding whether to hike or not. Hence,* $\mathcal{C} = \{c_0, c_1\}$ *where* c_0 *stands for not hiking and* c_1 *for hiking. The decision is based on four binary features: Being in vacation (*V*), having a concert (*C*), having a meeting (*M*) and having an exhibition (*E*), thus* $\mathcal{F} = \{V, C, M, E\}$ *and* $\mathtt{dom}(.) = \{0, 1\}$*. Assume a classification model* $\mathtt{F}$ *that assigns classes to instances of* $\mathcal{Y} = \{x_i \mid i = 1, \ldots, 7\} \subset \mathtt{Inst}$ *as shown in the table below.*

$\mathcal{Y}$	V	C	M	E	$\mathtt{F}(x_i)$
x_1	*0*	*0*	*1*	*0*	c_0
x_2	*1*	*0*	*0*	*0*	c_1
x_3	*0*	*0*	*1*	*1*	c_0
x_4	*1*	*0*	*0*	*1*	c_1
x_5	*0*	*1*	*1*	*0*	c_0
x_6	*0*	*1*	*1*	*1*	c_0
x_7	*1*	*1*	*0*	*1*	c_1

The set of literals $H = \{(V, 0), (M, 0)\}$ *is consistent. Consider the instance* x_1*, note that* $x_{1\downarrow H} = \{(V, 0),\ (M, 0),\ (C, 0),\ (E, 0)\}$*.*

3. Explanation Functions

Explaining a classifier means either describing its *global* behaviour, namely how it assigns classes in general and independently of instances, or *locally* justifying its predictions to individual instances. Whatever the goal, explanations may take different forms including natural language texts (eg. [35]), visualizations (eg. [36]), prototypes or examples (eg. [37]), conversations (eg. [38]), and attributes-values (eg. [8]). In this paper, we focus on the latter type where **an explanation is a set of literals**. We denote by $\mathcal{E}$ the set of all possible explanations, i.e., the set of all subsets of literals. Throughout this section, we consider a fixed but arbitrary theory $\mathtt{T} = \langle \mathcal{F}, \mathtt{dom}, \mathcal{C} \rangle$ and a classifier $\mathtt{F}$ which has predicted the outcomes of instances in $\mathtt{Inst}$.

Global explanations concern classes (i.e., predictions) and answer two questions: "why a given class is proposed?" or "why a class is not proposed?". A *class explainer* is a function which assigns to every class a set of explanations.

Definition 4 (Class Explainer). *A* class explainer *is a function* $\mathtt{G}$ *mapping every class* $c \in \mathcal{C}$ *into a set of explanations (i.e., a subset of* $\mathcal{E}$*).*

Local explanations concern instances (i.e., input data) and look for reasons behind their predictions. They may answer different questions like: "why a given class is proposed for the instance x?", or "why the prediction of the instance x has not been avoided?" or even "why the prediction of x is the class c instead of c'?". In this case, an explainer, called *instance explainer*, is a function which assigns to every instance a set of explanations.

Definition 5 (Instance Explainer). *An* instance explainer *is a function* $\mathtt{L}$ *mapping every instance* $x \in \mathtt{Inst}$ *into a set of explanations (i.e., a subset of* $\mathcal{E}$*).*

The two types of explainers are somehow related. Indeed, explaining for example why an instance received some class consists mainly of providing reasons behind assigning in general that class by the classifier. This means that an instance explanation cannot be different from the prediction rules of the classifier. We introduce below a notion of compatibility relating the two types of explainers. It states that the explanations of a class are nothing more than the explanations of instances labelled by that class.

Definition 6 (Compatibility). *Let* $\mathtt{T} = \langle \mathcal{F}, \mathtt{dom}, \mathcal{C} \rangle$ *be a theory,* $\mathtt{F}$ *a classifier,* $\mathtt{L}$ *an instance explainer and* $\mathtt{G}$ *a class explainer. We say that* $\mathtt{L}$ *and* $\mathtt{G}$ *are* compatible *iff, for every class* $c \in \mathcal{C}$, $\mathtt{G}(c) = \bigcup_{x \in \mathtt{Inst} \ s.t. \ \mathtt{F}(x)=c} \mathtt{L}(x)$.

We discuss below two other properties that any class/instance explainer would satisfy. Such properties are important for understanding the behaviour of an explainer, assessing its quality and for comparing pairs of explainers. The first property ensures the existence of explanations. This property may be mandatory for some types of explanations like *sufficient reasons* and not for others like *counterfactual*s. Sufficient reasons provide the main evidence behind assigning a class to an instance. Such explanations are required and seen as crucial feedback by the users of the classifier.

A counterfactual is a (minimal) change in instance for getting another outcome, which is better than the current one from a user's point of view. Such explanation may not exist if the explainer returns only counterfactuals whose changes are possible. Assume that for getting a bank loan, the age of a customer should be at most 55. Sending such information to a 60 year old customer is not necessary since the latter cannot modify his age.

Definition 7 (Success). *A class explainer* $\mathtt{G}$ *(resp. instance explainer* $\mathtt{L}$*) satisfies* success *iff for any class* $c \in \mathcal{C}$ *(resp. any instance* $x \in \mathtt{Inst}$*),* $\mathtt{G}(c) \neq \emptyset$ *(resp.* $\mathtt{L}(x) \neq \emptyset$*).*

The second property, called *coherence*, states that a set of literals cannot lead to two distinct predictions. As we will see later, this property is crucial for some type of explanations, namely sufficient reasons, called abductive explanations in the literature. Let us illustrate the idea with examples.

Example 1 (Cont.) Let us recall the rules of the classifier.

- If $E = 1$ and $F = 0$, then `Admitted`,
- If $E = 1$, $F = 1$, $W = 1$, then `Admitted`,
- If $E = 1$, $F = 1$ and $W = 0$ and $G = 1$, then `Admitted`,
- If $E = 1$, $F = 1$ and $W = 0$ and $G = 0$, then `Denied`,
- If $E = 0$, then `Denied`.

Coherence prevents an explainer from providing the reason $H = \{(E, 1), (F, 1), (W, 0)\}$ to the class `Admitted` and $H' = \{(E, 1), (F, 1), (W, 0), (G, 0)\}$ to the class `Denied`. In the example, H is incomplete and as such it does not explain properly the class.

Let us consider another example of situation that is prevented by the coherence property.

Example 2 (Cont.) Assume an explainer which provides sufficient reasons for predictions. Assume also that it explains the classes c_0 and c_1 with the sets $\{(V, 0)\}$ and $\{(M, 0)\}$ respectively. Note that the set $\{(V, 0), (M, 0)\}$ is consistent, then there exists at least one instance x in the feature space such that $\{(V, 0), (M, 0)\} \subseteq x$ (see Property 1). Hence $\{(V, 0)\}$ and $\{(M, 0)\}$ cannot be reasons for predicting c_0 and c_1 respectively.

To sum up, there are two undesirable situations that are prevented by coherence for a class explainer $\texttt{G}$. Let $c, c' \in \mathcal{C}$ such that $c \neq c'$.

1. $H \in \texttt{G}(c)$, $H' \in \texttt{G}(c')$ with $H \subseteq H'$.
2. $H \in \texttt{G}(c)$, $H' \in \texttt{G}(c')$ with $H \not\subseteq H'$ and $H \cup H'$ is consistent.

Definition 8 (Coherence). *A class explainer* $\texttt{G}$ *satisfies* coherence *iff for any two classes* $c, c' \in \mathcal{C}$ *such that* $c \neq c'$ *the following holds:* $\forall H \in \texttt{G}(c)$, $\forall H' \in \texttt{G}(c')$, $H \cup H'$ *is inconsistent.*

An instance explainer $\texttt{L}$ *satisfies* coherence *iff for any two instances* $x, x' \in \texttt{Inst}$ *such that* $\texttt{F}(x) \neq \texttt{F}(x')$ *the following holds:* $\forall H \in \texttt{L}(x)$, $\forall H' \in \texttt{L}(x')$, $H \cup H'$ *is inconsistent.*

The following result summarizes the links between the above properties.

Property 3. *Let* $\texttt{G}$ *and* $\texttt{L}$ *be a class explainer and an instance explainer respectively.*

- *If* $\texttt{G}$ *and* $\texttt{L}$ *are compatible and* $\texttt{G}$ *is coherent, then* $\texttt{L}$ *is coherent.*
- *If* $\texttt{G}$ *and* $\texttt{L}$ *are compatible and* $\texttt{L}$ *is coherent, then* $\texttt{G}$ *is coherent.*

4. Explaining Classifiers under Complete Information

Two approaches for explaining black-box classification models have been distinguished in the literature: a *global* approach which aims at stressing when classes are predicted independently of instances, and a *local* approach which looks for justifying individual predictions. In addition, different types of local explanations have been studied in the recent literature, however their links to global explanations remain unclear. In this section, we propose a unified setting for global explanations and local ones. It is based on dual concepts that provide global explanations: arguments in favour of predictions and arguments against predictions. The former justify why a class is suggested by a black-box classifier and the latter state why a class is not proposed. We investigate the properties of both types of arguments, and provide ways for generating arguments pro a class from arguments con the class and vice versa. Finally, we define abductive explanations by arguments pros and counterfactuals and contrastive explanations with arguments cons. The three types are defined under complete information since they are generated from the whole feature space. Throughout the section, we assume an arbitrary but fixed theory $\texttt{T} = \langle \mathcal{F}, \texttt{dom}, \mathcal{C} \rangle$ and an arbitrary but fixed classifier $\texttt{F}$.

4.1. Global Explanations

This section aims at understanding how a classifier $\mathtt{F}$ assigns classes to instances of a theory $\mathtt{T}$. For that purpose, we assume the availability of an oracle of $\mathtt{F}$ that can be queried on any instance. We are interested by the question: what is an argument in favour of labelling an instance with a class c? In what follows, we consider an argument as a set of literals that are minimally sufficient for labelling an instance with c. In other words, it is the smallest set of literals that always lead to the class c. Such arguments provide sufficient reasons for proposing a class c.

Definition 9 (Argument Pro). *An* argument pro *a class $c \in \mathcal{C}$ is a pair $\langle H, c\rangle$ such that:*

- $H \subseteq \mathtt{Lit}$
- *H is consistent*
- $\forall x \in \mathtt{Inst}$ *such that* $H \subseteq x$, $\mathtt{F}(x) = c$
- *$\nexists H' \subset H$ such that H' satisfies the third condition.*

H and c are respectively called support *and* conclusion *of the argument. Let $\mathtt{Pros}(c)$ denote the set of all arguments pro c in theory $\mathtt{T}$, and $\mathtt{arg}^{+}(\mathtt{T}) = \bigcup\limits_{c\in\mathcal{C}} \mathtt{Pros}(c)$, i.e., $\mathtt{arg}^{+}(\mathtt{T})$ stands for the set of all arguments pro classes of a theory. Let $\mathtt{G}_{pro}$ be the class explainer which assigns to every class $c \in \mathcal{C}$ a set $\{H \mid \langle H, c\rangle \in \mathtt{Pros}(c)\}$.*

The consistency condition is useful for discarding irrelevant arguments, thus global explanations, of the form $\langle\{(f_1, v_1),\ (f_1, v_2)\},\ c\rangle$.

Example 1 (Cont.) It can be checked that $\mathtt{G}_{pro}(\mathtt{Denied}) = \{H_1, H_2\}$ such that:

- $H_1 = \{(E, 0)\}$,
- $H_2 = \{(E, 1), (F, 1), (W, 0), (G, 0)\}$.

$\mathtt{G}_{pro}(\mathtt{Amitted}) = \{H'_1, H'_2, H'_3\}$ such that:

- $H'_1 = \{(E, 1), (F, 0)\}$,
- $H'_2 = \{(E, 1), (F, 1), (W, 1)\}$,
- $H'_3 = \{(E, 1), (F, 1), (W, 0), (G, 1)\}$.

Example 3. *Let* $\mathtt{T} = \langle \mathcal{F},\ \mathtt{dom},\ \mathcal{C} \rangle$ *be such that* $\mathcal{F} = \{f_1, f_2\}$, $\mathtt{dom}(f_1) = \mathtt{dom}(f_2) = \{0, 1\}$, *and* $\mathcal{C} = \{c_1, c_2, c_3\}$. *Assume the following assignments of classes to instances by a classifier* $\mathtt{F}$.

Inst	f_1	f_2	$\mathtt{F}(x_i)$
x_1	*0*	*0*	c_1
x_2	*0*	*1*	c_2
x_3	*1*	*0*	c_3
x_4	*1*	*1*	c_3

The classes c_1, c_2, c_3 *are supported respectively by the arguments* a_1, a_2, a_3 *such that:*

- $a_1 = \langle \{(f_1, 0), (f_2, 0)\}, c_1 \rangle$ $\qquad \mathtt{G}_{pro}(c_1) = \{\{(f_1, 0), (f_2, 0)\}\}$
- $a_2 = \langle \{(f_1, 0), (f_2, 1)\}, c_2 \rangle$ $\qquad \mathtt{G}_{pro}(c_2) = \{\{(f_1, 0), (f_2, 1)\}\}$
- $a_3 = \langle \{(f_1, 1)\}, c_3 \rangle$ $\qquad \mathtt{G}_{pro}(c_3) = \{\{(f_1, 1)\}\}$

A class may have zero, one, or several arguments pro. The first case holds when the class is not assigned by the classification model $\mathtt{F}$ to any instance. When the same class is assigned to all instances of $\mathtt{Inst}$, then the set of arguments would contain a single argument, which is in favour of the class and its support is the empty set. Furthermore, from a theory, it is possible to generate arguments in favour of any class provided that the latter is ascribed to at least one instance. Finally, every argument refers to at least one instance of $\mathtt{Inst}$. Note that from the same instance, it is possible to generate more than one argument in favour of a class.

Proposition 1. *Let* $c \in \mathcal{C}$. *The following properties hold:*

1. $(\mathtt{arg}^+(\mathtt{T}) = \{\langle \emptyset, c \rangle\}) \iff (\forall x \in \mathtt{Inst},\ \mathtt{F}(x) = c)$
2. *For any* $x \in \mathtt{Inst}$, *if* $\mathtt{F}(x) = c$, *then* $\exists \langle H, c \rangle \in \mathtt{Pros}(c)$. *Furthermore,* $H \subseteq x$.
3. *If* $\langle H, c \rangle \in \mathtt{Pros}(c)$, *then* $\exists x \in \mathtt{Inst}$ *such that* $\mathtt{F}(x) = c$.
4. $\mathtt{Pros}(c) = \emptyset$ *iff* $\forall x \in \mathtt{Inst},\ \mathtt{F}(x) \neq c$.

The fourth property in the previous result shows that the only case where a class is not supported by argument is when the class is never assigned by the classification model. This case is extreme as it does not occur in practice. Hence, if a classifier is surjective, then pros exist for every class and consequently, $\mathtt{G}_{pro}$ satisfies success.

Proposition 2. *If the classifier* $\mathtt{F}$ *is a surjective function, then* $\mathtt{G}_{pro}$ *satisfies success.*

The following result shows that the supports of any pair of arguments pro distinct classes are inconsistent. This means that the class explainer $\mathtt{G}_{pro}$ satisfies coherence.

Proposition 3. *Let* $c, c' \in \mathcal{C}$ *with* $c \neq c'$*. For all* $\langle H, c\rangle, \langle H', c'\rangle \in \mathtt{arg}^+(\mathtt{T})$*, the set* $H \cup H'$ *is inconsistent. And,* $\mathtt{G}_{pro}$ *satisfies coherence.*

We show next that the arguments that can be generated from a theory define a partition of the set $\mathtt{Inst}$ of instances.

Proposition 4. *Let* $\mathcal{C} = \{c_1, \ldots, c_m\}$ *and* $i \in \{1, \ldots, m\}$*,*

$$\mathtt{Inst}_i = \{x \in \mathtt{Inst} \mid \exists \langle H, c_i\rangle \in \mathtt{arg}^+(\mathtt{T}) \text{ and } H \subseteq x\}.$$

The following properties hold:

1. *For all* $i, j \in \{1, \ldots, m\}$ *such that* $i \neq j$*,* $\mathtt{Inst}_i \cap \mathtt{Inst}_j = \emptyset$*.*
2. $\mathtt{Inst} = \mathtt{Inst}_1 \cup \ldots \cup \mathtt{Inst}_m$*.*

We now introduce the notion of argument *against* or *con* a class, say c. It is a minimal set of literals that is sufficient for not assigning the class c to any instance. It defines explanations that answer the question “why c is not recommended by a classifier?” or “why not c?”.

Notation: For $c \in \mathcal{C}$, $\overline{c}$ denotes that c is not recommended.

Definition 10 (Argument Con). *Let* $c \in \mathcal{C}$*. An* argument con c *is a pair* $\langle H, \overline{c}\rangle$ *such that:*

- $H \subseteq \mathtt{Lit}$
- H *is consistent*
- $\forall x \in \mathtt{Inst}$ *such that* $H \subseteq x$*,* $\mathtt{F}(x) \neq c$
- $\nexists H' \subset H$ *such that* H' *satisfies the third condition.*

Let $\mathtt{Cons}(c)$ *be the set of all arguments con* c *and* $\mathtt{arg}^-(\mathtt{T}) = \bigcup\limits_{c \in \mathcal{C}} \mathtt{Cons}(c)$*, i.e.,* $\mathtt{arg}^-(\mathtt{T})$ *stands for the set of all arguments con classes of a theory. Let* $\mathtt{G}_{con}$ *be the class explainer which assigns to every class* $c \in \mathcal{C}$ *a set* $\{H \mid \langle H, \overline{c}\rangle \in \mathtt{Cons}(c)\}$*.*

Example 3 (Cont.) The classes c_1, c_2, c_3 have the following arguments con.

- $b_1 = \langle \{(f_1, 1)\}, \overline{c_1} \rangle$ $\qquad$ $\mathtt{G}_{con}(c_1) = \{\{(f_1, 1)\}, \{(f_2, 1)\}\}$
- $b_2 = \langle \{(f_2, 1)\}, \overline{c_1} \rangle$
- $b_3 = \langle \{(f_1, 1)\}, \overline{c_2} \rangle$ $\qquad$ $\mathtt{G}_{con}(c_2) = \{\{(f_1, 1)\}, \{(f_2, 0)\}\}$
- $b_4 = \langle \{(f_2, 0)\}, \overline{c_2} \rangle$
- $b_5 = \langle \{f_1, 0\}, \overline{c_3} \rangle$ $\qquad$ $\mathtt{G}_{con}(c_3) = \{\{f_1, 0\}\}$

This example shows clearly that coherence is not satisfied by the class explainer $\mathtt{G}_{con}$. It is even not recommended when explaining why a class does not hold. Indeed, two classes like c_1 and c_2 may have the same reason for avoiding them, namely $\{(f_1, 1)\}$. This shows a key difference between explaining "why a class is suggested" and "why it is not suggested".

Proposition 5. *The class explainer* $\mathtt{G}_{con}$ *violates coherence.*

It is easy to show that when the concept to learn is binary, the arguments pro one class are con the other.

Proposition 6. *If* $\mathcal{C} = \{c, c'\}$*, then:*

1. $\mathtt{Pros}(c) = \{\langle H, c \rangle \mid \langle H, \overline{c'} \rangle \in \mathtt{Cons}(c')\}$,
2. $\mathtt{Cons}(c) = \{\langle H, \overline{c} \rangle \mid \langle H, c' \rangle \in \mathtt{Pros}(c')\}$.

In case of non-binary concepts, an argument that is against a given class does not necessarily support another class. Let us consider the following abstract example.

Example 3 (Cont.) The argument $\langle \{(f_1, 0)\}, \overline{c_3} \rangle$ is against c_3, however the set $\{(f_1, 0)\}$ is not sufficient for supporting any other class.

Naturally, the support of every argument against a class is inconsistent with the support of any argument pro that class.

Proposition 7. *Let* $c \in \mathcal{C}$*. For all* $\langle H, c \rangle \in \mathtt{Pros}(c)$*,* $\langle H', \overline{c} \rangle \in \mathtt{Cons}(c)$*, the set* $H \cup H'$ *is inconsistent.*

The following results show the relationship between an argument against a class and those supporting other classes.

Proposition 8. *Let $c \in \mathcal{C}$. The following properties hold:*

1. $\langle \emptyset, \overline{c} \rangle \in \mathtt{Cons}(c)$ *iff* $\forall x \in \mathtt{Inst}, \mathtt{F}(x) \neq c$.
2. *If* $\langle H, \overline{c} \rangle \in \mathtt{Cons}(c)$*, then* $\exists x \in \mathtt{Inst}$ *such that* $\mathtt{F}(x) \neq c$*. Furthermore,* $H \subseteq x$*.*
3. *If* $\exists x \in \mathtt{Inst}$ *such that* $\mathtt{F}(x) \neq c$*, then* $\exists \langle H, \overline{c} \rangle \in \mathtt{Cons}(c)$ *such that* $H \subseteq x$*.*
4. *If* $\langle H, c \rangle \in \mathtt{Pros}(c)$*, then* $\forall c' \in \mathcal{C} \setminus \{c\}$*,* $\exists \langle H', \overline{c'} \rangle \in \mathtt{Cons}(c')$ *such that* $H' \subseteq H$*.*

While a class that is not assigned to any instance has no pros, we show that it has a single argument con whose support is the empty set.

Proposition 9. *Let $c \in \mathcal{C}$. The following equivalences hold:*

1. $(\mathtt{Pros}(c) = \emptyset) \iff (\mathtt{Cons}(c) = \{\langle \emptyset, \overline{c} \rangle\})$
2. $(\mathtt{Cons}(c) = \emptyset) \iff (\mathtt{Pros}(c) = \{\langle \emptyset, c \rangle\})$

We have also the following straightforward property.

Proposition 10. *Let $c \in \mathcal{C}$. It holds that* $\mathtt{Inst} = \mathcal{Y} \cup \mathcal{Z}$ *where*

$$\mathcal{Y} = \{x \in \mathtt{Inst} \mid \exists \langle H, \overline{c} \rangle \in \mathtt{Cons}(c) \text{ and } H \subseteq x\},$$

$$\mathcal{Z} = \{x \in \mathtt{Inst} \mid \exists \langle H, c \rangle \in \mathtt{Pros}(c) \text{ and } H \subseteq x\}.$$

From this property, it follows that if the classifier is a surjective function, then the class explainer $\mathtt{G}_{con}$ satisfies success.

Proposition 11. *If the classifier* $\mathtt{F}$ *is a surjective function, then* $\mathtt{G}_{con}$ *satisfies success.*

We show next that the two class explainers provide *dual* explanations. Indeed, the sufficient reasons for proposing a class c can be generated from the sufficient reasons for discarding the class and vice versa. The first result below shows how to define the elements of the set $\mathtt{G}_{pro}(c)$ from those of the set $\mathtt{G}_{con}$.

Theorem 1. *Let $c \in \mathcal{C}$ and $H \subseteq \mathtt{Lit}$. $H \in \mathtt{G}_{pro}(c)$ iff the following hold:*

- *H is consistent*
- *$\forall H' \in \mathtt{G}_{con}(c)$, $H \cup H'$ is inconsistent,*

- *$\nexists H'' \subseteq$ `Lit` such that $H'' \subset H$ and H'' satisfies the second condition.*

The following result shows how to generate sufficient reasons for discarding a class from sufficient reasons for proposing the class to instances.

Theorem 2. *Let $c \in \mathcal{C}$ and $H \subseteq$ `Lit`. $H \in \mathtt{G}_{con}(c)$ iff the following hold:*

- *H is consistent*
- *$\forall H' \in \mathtt{G}_{pro}(c)$, $H \cup H'$ is inconsistent,*
- *$\nexists H'' \subseteq$ `Lit` such that $H'' \subset H$ and H'' satisfies the second condition.*

To sum up, this section discussed two types of global explanations under complete information: sufficient reasons for assigning a class to instances and sufficient reasons for not suggesting the class. We have shown that the two types are dual, and thus each of them can be generated from the other. Both types of explanations exist when the classifier is a surjective function. Note that in practice, this constraint is satisfied since every class is assigned to at least one instance. The two types are however distinguished by the coherence property. We have seen that this property is mandatory for explaining why a class is suggested otherwise incorrect answers could be provided by the corresponding class explainer. However, coherence is not required for answering why a class is not proposed.

4.2. Local Explanations

This section investigates three types of local explanations: abductive explanations, contrastive explanations and counterfactual explanations. It studies their formal properties and their links with the global explanations which are input-dependent. We show that local explanations, whatever their type, are generated from arguments pro/con classes.

4.2.1. Abductive Explanations

Abductive explanations answer the question: "why an instance x is labelled with a class c"?, i.e., why does the outcome c hold for x? The answer consists in highlighting *factors that determined the given class.* In [11, 12, 18, 39, 40], an abductive explanation, called also prime implicant in [11, 18], is defined as a *minimal* (for set inclusion) set of literals that is sufficient for predicting a class. It is thus a sufficient reason for assigning a class to an instance. Such explanations are closely tied to arguments pro classes. They are definitely the supports of arguments pro classes. In what follows, we refer to them as absolute explanations since they are generated from the whole feature space (see condition 3 of Definition 9)

Definition 11 (Absolute Abductive Explanation). *Let $x \in \mathtt{Inst}$. An* absolute abductive explanation *of x is any member of the set:*

$$\mathtt{L}_{ae}(x) = \{H \subseteq \mathtt{Lit} \mid H \in \mathtt{G}_{pro}(c) \text{ and } H \subseteq x\}.$$

Example 1 (Cont.) Assume a candidate who has the following profile: $\{(E,0),(F,),(W,0),(G,0)\}$. The Bayesian classifier assigns the class `Denied`. There is a single explanation in this case: $\mathtt{L}_{ae}(x) = \{H_1\}$ where $H_1 = \{(E,0)\}$. Recall that the class `Denied` has another global explanation, which is $H_2 = \{(E,1),(F,1),(W,0),(G,0)\}$.

Example 3 (Cont.) The absolute abductive explanations of x_1, x_2, x_3, x_4 are as follows:

- $\mathtt{L}_{ae}(x_1) = \{H_1\}$ $\qquad H_1 = \{(f_1,0),(f_2,0)\}$
- $\mathtt{L}_{ae}(x_2) = \{H_2\}$ $\qquad H_2 = \{(f_1,0),(f_2,1)\}$
- $\mathtt{L}_{ae}(x_3) = \{H_3\}$ $\qquad H_3 = \{(f_1,1)\}$
- $\mathtt{L}_{ae}(x_4) = \{H_3\}$

From the results presented in the previous section, it follows that a class that is assigned to all instances has a unique explanation, which is the emptyset.

Proposition 12. *Let $x \in \mathtt{Inst}$.*

1. $\mathtt{L}_{ae}(x) = \{\emptyset\}$ *iff* $\forall y \in \mathtt{Inst}$, $\mathtt{F}(y) = \mathtt{F}(x)$.
2. $\mathtt{L}_{ae}(x) \subseteq \mathtt{G}_{pro}(c)$

We show that the absolute abductive explanation function is coherent and satisfies success. Furthermore, it is compatible with $\mathtt{G}_{pro}$.

Proposition 13. *The function $\mathtt{L}_{ae}$ satisfies coherence and success. Furthermore, $\mathtt{L}_{ae}$ and $\mathtt{G}_{pro}$ are compatible.*

Remark: It is worth mentioning that in [15], an abductive explanation of an instance x is defined as the minimal set of *features* (instead of literals) that determine the class $\mathtt{F}(x)$. Formally, the set of explanations of x is: $\{\{f \in \mathcal{F} \mid (f,t) \in H\}$ where $H \in \mathtt{L}_{ae}(x)\}$. This feature-based definition is reasonable when providing local explanations (i.e., for instances), however unlike our definition, it **may not recover the global explanations** of a class as shown in the following example.

Example 3 (Cont.) The feature f_1 determines the prediction c_3, hence $\{f_1\}$ is an explanation of x_3 and x_4. However, this is true only when f_1 gets the value 1. Indeed, if f_1 receives 0, then c_3 is not recommended by the classifier. Thus, from the explanations of the instances, it is not possible to deduce those of the classes.

4.2.2. Counterfactual Explanations

Counterfactual explanations are widely used for interpreting predictions of black-box machine learning models, see eg. [23, 24, 41]. In the literature, such explanations are sometimes confused with contrastive explanations. In [6] a clear distinction is made between the two notions. Counterfactuals state how the outcome of a given instance could have been changed. For that purpose, they provide the (minimal) change in an instance that is sufficient for altering the prediction of the instance to *whatever class*. The key idea is then to avoid the current class of an instance, and this is exactly what arguments cons a class provide.

Definition 12 (Counterfactual Explanation). *Let $x \in \texttt{Inst}$. The* counterfactual explanations *of x are the minimal (for set inclusion) elements of the set:*

$$\{H \setminus x \mid \langle H, \overline{\texttt{F}(x)}\rangle \in \texttt{Cons}(\texttt{F}(x))\}.$$

We denote by $\texttt{L}_{cf}(x)$ the set of all counterfactual explanations of x.

Example 4. *Consider the theory below:*

Inst	f_1	f_2	f_3	$\texttt{F}(x_i)$
x_1	0	0	0	c_1
x_2	0	0	1	c_1
x_3	0	1	0	c_1
x_4	0	1	1	c_2
x_5	1	0	0	c_1
x_6	1	0	1	c_3
x_7	1	1	0	c_3
x_8	1	1	1	c_3

Let us focus on the instance x_2 and how to avoid its outcome c_1. The class c_1 has three arguments con $\langle U_i, \overline{c_1}\rangle$ each of which leads to some H_i:

- $U_1 = \{(f_1, 1), (f_2, 1)\}$ $H_1 = U_1 \setminus x_2 = \{(f_1, 1), (f_2, 1)\}$
- $U_2 = \{(f_1, 1), (f_3, 1)\}$ $H_2 = U_2 \setminus x_2 = \{(f_1, 1)\}$

- $U_3 = \{(f_2, 1), (f_3, 1)\}$ $\qquad H_3 = U_3 \setminus x_2 = \{(f_2, 1)\}$

The counterfactual explanations of x_2 are H_2 and H_3. They state respectively that the value of f_1 and f_2 should be modified in order to avoid the class current c_1.

Remark: Note that arguments con a class may intersect with instances to which the class is assigned (e.g. U_2 and U_3 share the literal $(f_3, 1)$ with x_2). The definition of counterfactual explanation removes those common literals.

The following result provides a characterization of counterfactual explanations.

Theorem 3. *For any $x \in \mathtt{Inst}$, $H \in \mathtt{L}_{cf}(x)$ iff H satisfies the conditions below:*

- $H \subseteq \mathtt{Lit}$
- *H is consistent*
- $\mathtt{F}(x_{\downarrow H}) \neq \mathtt{F}(x)$
- *$\nexists H' \subset H$ such that H' satisfies the above conditions.*

Remark: Our definition of counterfactual explanation generalizes and solves a drawback of the notion of contrastive explanation as presented in [15]. In that paper, a contrastive explanation for an instance x is defined as a minimal (for set inclusion) subset of features $F \subseteq \mathcal{F}$ such that $\exists y \in \mathtt{Inst} \setminus \{x\}$ where $\mathtt{F}(y) \neq \mathtt{F}(x)$ and $\forall f \in \mathcal{F} \setminus F$, $\mathtt{Val}(f, x) = \mathtt{Val}(f, y)$. In Example 4, the contrastive explanations of x_2 are $\{f_1\}$ and $\{f_2\}$. Such a definition is reasonable when all features are binary since the modified value of each attribute is implicit. This is however not true in the general case as shown in the following example.

Example 5. *Consider the following theory and a classifier that provides the predictions described in the table below.*

Inst	f_1	f_2	$\mathtt{F}(x_i)$
x_1	*0*	*0*	c_1
x_2	*0*	*1*	c_2
x_3	*0*	*2*	c_1
x_4	*1*	*0*	c_3
x_5	*1*	*1*	c_3
x_6	*1*	*2*	c_3

The instance x_1 has two contrastive explanations in the sense of [15]: $\{f_1\}$ and $\{f_2\}$. However, it is worth noticing that f_2 can only be modified to 1 *since when it receives the value 2, the prediction does not change.*

Unlike abductive explanations, counterfactual explanations may not exist. This is particularly the case when the class of the input at hand is assigned to all instances of the feature space. However, they can never be the empty set.

Proposition 14. *Let $x \in \texttt{Inst}$. The following hold:*

1. $\texttt{L}_{cf}(x) \subseteq \{H \setminus x \mid H \in \texttt{G}_{con}(c)\}$,
2. $\texttt{L}_{cf}(x) = \emptyset$ *iff* $\forall y \in \texttt{Inst}$, $\texttt{F}(y) = \texttt{F}(x)$,
3. $\emptyset \notin \texttt{L}_{cf}(x)$

We have seen in the previous section that the property of coherence is not required for the class explainer $\texttt{G}_{con}$. It is also not required for counterfactual explanations since the same minimal change may be necessary for two instances which have different predictions.

Proposition 15. *The instance explainer $\texttt{L}_{cf}$ violates coherence. If the classifier $\texttt{F}$ is a surjective function, then $\texttt{L}_{cf}$ satisfies success.*

4.2.3. Contrastive Explanations

A predominant finding from research in the philosophy of science and social sciences is that explanation-seeking behaviour is generally contrastive [42]. Indeed, when asking why a model predicted an output, humans ask for a contrast against an *expected* output, called *foil* in the literature. They answer thus the question: *Why class c rather than class c′?*. Contrastive explanations oppose thus the current prediction of an instance to another (generally desirable) outcome [10, 6]. In what follows, we define a contrastive explanation as a minimal change in the instance which leads to the expected class.

Definition 13 (Contrastive Explanation). *Let $x \in \texttt{Inst}$, $c \in \mathcal{C}$ such that $\texttt{F}(x) \neq c$. A* contrastive explanation *of (x, c) is a set $H \subseteq \texttt{Lit}$ such that:*

- *H is consistent,*
- *$\exists y \in \texttt{Inst}$ such that $\texttt{F}(y) = c$ and $y = x_{\downarrow H}$,*

- *$\nexists H' \subset H$ s.t. H' satisfies the above conditions.*

Let $\mathtt{L}_{co}(x, c)$ denote the set of all contrastive explanations of (x, c).

Example 3 (Cont.) It is easy to check that $\mathtt{L}_{co}(x_1, c_2) = \{\{(f_2, 1)\}\}$. The informal reading: *if the feature f_2 takes the value 1, then the outcome will no longer be c_1 but rather c_2*. Note also that $\mathtt{L}_{co}(x_1, c_3) = \{\{(f_1, 1)\}\}$.

In what follows, we show that contrastive explanations and counterfactual explanations may be different.

Example 6. *Consider the theory below:*

Inst	f_1	f_2	f_3	$\mathtt{F}(x_i)$
x_1	*1*	*0*	*0*	c_1
x_2	*0*	*1*	*1*	c_2
x_3	*1*	*0*	*1*	c_3
x_4	*1*	*1*	*0*	c_3
x_5	*1*	*1*	*1*	c_3

Let us focus on the instance x_1 and how to avoid its outcome c_1. It can be checked that x_1 has two counterfactual explanations: $H_1 = \{(f_2, 1)\}$ and $H_2 = \{(f_3, 1)\}$. Assume now that our expected outcome is the class c_2. Note that none of the two possible changes lead to the desirable outcome. Indeed, $x_{1\downarrow H_1} = x_4$ and $\mathtt{F}(x_4) \neq c_2$ and $x_{1\downarrow H_2} = x_4$ and $\mathtt{F}(x_3) \neq c_2$. The pair (x_1, c_2) has a single contrastive explanation which is $\{(f_1, 0), (f_2, 1), (f_3, 1)\}$.

When the concept to learn is binary, contrastive explanations coincide with the counterfactual ones. This is not surprising since there is only one possible targeted class.

Proposition 16. *Let $\mathtt{T} = \langle \mathcal{F}, \mathtt{dom}, \mathcal{C} \rangle$ be a theory, $\mathcal{C} = \{c, c'\}$, and $x \in \mathtt{Inst}$ such that $\mathtt{F}(x) = c$. It holds that $\mathtt{L}_{co}(x, c') = \mathtt{L}_{cf}(x)$.*

We show next that the function $\mathtt{L}_{co}$ satisfies success when the classifier is surjective, but it violates coherence.

Proposition 17. *The instance explainer $\mathtt{L}_{co}$ satisfies success when the classifier is surjective. It violates coherence.*

5. Explaining Classifiers under Incomplete Information

All types of explanations that we introduced in the previous section require exploring the whole feature space. While the corresponding explainers satisfy interesting properties, the approach may not be feasible, in particular for complex classifiers whose querying may not be reasonable for all instances. In this section, we investigate explanations under incomplete information, i.e., only a subset of instances is considered. The latter may be the dataset a classifier has been trained on, a dataset in which the classifier has better performance, etc. It is worth mentioning that some well-known explanation functions like Anchors [9] and LIME [8]) already follow this approach and use datasets that they generate in specific ways. Throughout this section, we focus on local explanations, namely sufficient reasons for instance prediction, or abductive explanations. We consider an arbitrary theory $\mathtt{T} = \langle \mathcal{F}, \mathtt{dom}, \mathcal{C} \rangle$, a subset $\mathcal{Y} \subseteq \mathtt{Inst}$ of instances and a classifier $\mathtt{F}$.

5.1. Plausible Explanations

We define an explanation function that provides minimally sufficient reasons from a dataset $\mathcal{Y}$. Such a definition abstracts Anchors and LIME since they both use datasets generated in different ways. Explanations are based on arguments pro classes generated from $\mathcal{Y}$ as follows.

Definition 14 (Argument). *Let $c \in \mathcal{C}$. An* argument *in favor of c is a pair $\langle H, c\rangle$ where:*

- $H \subseteq \mathtt{Lit}$,
- $\exists x \in \mathcal{Y}$ *such that* $H \subseteq x$,
- $\forall y \in \mathcal{Y}$ *s.t.* $H \subseteq y$, $\mathtt{F}(y) = c$,
- $\nexists H' \subset H$ *that verifies the above conditions.*

H and c are called respectively support *and* conclusion *of the argument. Let* $\arg(\mathcal{Y})$ *denote the set of arguments built from $\mathcal{Y}$.*

Note that the set $\arg(\mathcal{Y})$ is finite since $\mathcal{Y}$ is finite. This set is used for defining plausible abductive explanations for instances.

Definition 15 (Plausible Explanation). *Let $x \in \mathtt{Inst}$. A* plausible explanation *of x is any member of the set:*

$$\mathtt{L}_{pe}(x) = \{H \subseteq \mathtt{Lit} \mid \langle H, \mathtt{F}(x)\rangle \in \arg(\mathcal{Y}) \textit{ and } H \subseteq x\}.$$

Consider the initial running example which contains seven instances.

Example 2 (Cont.) There are two classes in the theory: c_0, c_1. Their arguments are given below:

- $a_1 = \langle U_1, c_0 \rangle$ $\quad U_1 = \{(V, 0)\}$
- $a_2 = \langle U_2, c_0 \rangle$ $\quad U_2 = \{(M, 1)\}$
- $a_3 = \langle U_3, c_0 \rangle$ $\quad U_3 = \{(C, 1), (E, 0)\}$
- $a_4 = \langle U_4, c_1 \rangle$ $\quad U_4 = \{(V, 1)\}$
- $a_5 = \langle U_5, c_1 \rangle$ $\quad U_5 = \{(M, 0)\}$

It can be checked that:

- $\mathtt{L}_{pe}(x_1) = \{U_1, U_2\}$
- $\mathtt{L}_{pe}(x_5) = \{U_1, U_2, U_3\}$
- $\mathtt{L}_{pe}(x_2) = \mathtt{L}_{pe}(x_4) = \mathtt{L}_{pe}(x_7) = \{U_4, U_5\}$

Every plausible abductive explanation is consistent. Furthermore, an instance may have one or several (absolute, plausible) abductive explanations.

Proposition 18. *Let $x \in \mathcal{Y}$.*

1. *For any $H \in \mathtt{L}_{pe}(x)$, H is consistent.*
2. *If $\mathcal{Y} = \mathtt{Inst}$, then $\mathtt{L}_{ae}(x) = \mathtt{L}_{pe}(x)$*
3. *$\mathtt{L}_{pe}(x) = \{\emptyset\}$ iff $\forall y \in \mathcal{Y} \setminus \{x\}$, $\mathtt{F}(y) = \mathtt{F}(x)$.*

The plausible abductive explanation function $\mathtt{L}_{pe}$ guarantees at least one explanation for every instance, however it violates coherence. This means that it may return incorrect explanations. Indeed, in Example 2, the two instances x_1 and x_2 are assigned different classes. However, $\{(V, 0)\}$ is a plausible explanation of x_1 and $\{(M, 0)\}$ is a plausible explanation of x_2. Note that the set $\{(V, 0), (M, 0)\}$ is consistent. Thus, there exists for sure an instance, say $z \in \mathtt{Inst}$, such that $\{(V, 0), (M, 0)\} \subseteq z$. Hence, the third condition of Definition 15 would not be applicable, which means that at least one of two explanations is incorrect.

Proposition 19. *The function $\mathtt{L}_{pe}$ violates coherence and satisfies success.*

Remark: As noted earlier, Anchors and LIME explanation functions are somehow instances of $\mathtt{L}_{pe}$ as they generate abductive explanations from a proper subset of the feature space. They even do not use a whole dataset but only instances that are in the neighbourhood of the instance being explained. Hence, the two functions violate coherence.

Let us switch to another property of the instance explainer $\mathtt{L}_{pe}$. We show that a plausible explanation of an instance is not necessarily an absolute one, while, ideally $\mathtt{L}_{pe}$ should approximate $\mathtt{L}_{ae}$.

Property 4. *Let $\mathcal{Y} \subset \mathtt{Inst}$ and $x \in \mathcal{Y}$. $\mathtt{L}_{pe}(x) \not\subseteq \mathtt{L}_{ae}(x)$.*

Example 2 (Cont.) Consider the instance x_5. Recall that $U_3 = \{(C, 1), (E, 0)\}$ is a plausible explanation of x_5. Assume we receive the new instance x_8 below:

$\mathcal{Y}$	V	C	M	E	$\mathtt{F}(x_8)$
x_8	1	1	0	0	c_1

Note that $\{(C, 1), (E, 0)\}$ is no longer a plausible explanation of x_5 that can be generated from the set $\mathcal{Y} \cup \{x_5\}$.

We show next that when the set of plausible explanations generated by $\mathtt{L}_{pe}$ is incoherent, then any explainer which extracts a subset of those explanations cannot guarantee success and coherence together. Before presenting the formal result, let us first introduce plausible explainers.

Definition 16 (Plausible Explainer). *Let $\mathcal{Y} \subseteq \mathtt{Inst}$. A* plausible explainer *is a function $\mathtt{L}$ mapping every instance $x \in \mathcal{Y}$ into $\mathtt{L}(x) \subseteq \mathtt{L}_{pe}(x)$.*

We show next that there is no plausible explainer that can satisfy the two principles at the same time. This impossibility result is important as it shows the existence of a dilemma between two desirable properties.

Theorem 4. *There is no plausible explainer that satisfies both coherence and success.*

Depending on the application domain, one can choose the property to be satisfied. In critical applications like healthcare, providing correct explanations is crucial as they may be used for making further decisions like prescribing treatments for patients.

To sum up, explanations are constructed from a dataset, which is a subset of the feature space of a theory. However, due to incompleteness of information in the dataset, some explanations may be incorrect, i.e., they are not absolute. Furthermore, we have seen that the actual functions (like Anchors and LIME) that generate plausible explanations suffer from another weakness which is incoherence. The latter leads also to incorrect explanations. Thus, defining an explanation function that is coherent remains a challenge in the literature. However, the impossibility result shows that satisfaction of coherence would be at the expense of success. In the next sections, we propose two such functions.

5.2. Argument-based Explanation Function

We propose a novel explanation function, which is based on arguments (see Definition 14). The latter support classes, in the sense they provide the minimal sets of literals that determine a class. They are thus independent from instances. An advantage of not considering instances is to reduce the number of arguments that can be built. The arguments may be conflicting. This is particularly the case when they violate the coherence property, namely when their supports are consistent but their conclusions are different.

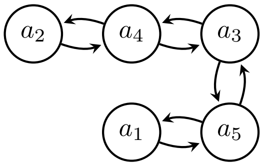
Definition 17 (Attack Relation). *Let $\langle H, c\rangle$, $\langle H', c'\rangle \in \arg(\mathcal{Y})$. We say that $\langle H, c\rangle$ attacks $\langle H', c'\rangle$ iff:*

- *$H \cup H'$ is consistent, and*
- *$c \neq c'$.*

Obviously, the above attack relation is symmetric and irreflexive.

Property 5. *Let $a, b \in \arg(\mathcal{Y})$. If a attacks b, then b attacks a. Furthermore, an argument does not attack itself.*

Example 2 (Cont.) The attacks between the arguments are depicted in the figure below:



In this example, every argument in favor of a class attacks at least one argument in favour of the other class. This shows that the plausible explanations generated by the function $\mathtt{L}_{pe}$ are incoherent, and cannot all be correct.

Arguments and their attack relationships form what we call an argumentation system.

Definition 18 (Argumentation System). *An* argumentation system *built from $\mathcal{Y} \subseteq$ `Inst` is a pair $AS = \langle \arg(\mathcal{Y}), \mathcal{R} \rangle$ where $\mathcal{R} \subseteq \arg(\mathcal{Y}) \times \arg(\mathcal{Y})$ such that for $a, b \in \arg(\mathcal{Y})$, $(a, b) \in \mathcal{R}$ iff a attacks b (in the sense of Definition 17).*

Since arguments are conflicting, they should be evaluated using a semantics. There are different types of semantics in the literature. In this paper, we consider extension-based ones that have been introduced by Dung in [33]. They compute sets of arguments that can be jointly accepted. Each set is called an extension and represents a coherent position. In this paper, we focus on stable semantics defined as follows.

Definition 19 (Stable Extensions). *Let $AS = \langle \arg(\mathcal{Y}), \mathcal{R} \rangle$ be an argumentation system and $\mathcal{E} \subseteq \arg(\mathcal{Y})$. The set $\mathcal{E}$ is a* stable extension *iff:*

- *$\nexists a, b \in \mathcal{E}$ such that $(a, b) \in \mathcal{R}$, and*
- *$\forall a \in \arg(\mathcal{Y}) \setminus \mathcal{E}$, $\exists b \in \mathcal{E}$ such that $(b, a) \in \mathcal{R}$.*

Let $\sigma(AS)$ denote the set of all stable extensions of AS.

It is well-known that stable extensions may not exist in some cases. However, since the attack relation is symmetric and irreflexive, stable extensions coincide with the naive ones which always exist in this case as shown in [43].

Example 2 (Cont.) The argumentation system has four stable extensions:

- $\mathcal{E}_1 = \{a_1, a_2, a_3\}$
- $\mathcal{E}_2 = \{a_1, a_4\}$
- $\mathcal{E}_3 = \{a_2, a_5\}$
- $\mathcal{E}_4 = \{a_4, a_5\}$

Each stable extension refers to a possible set of explanations. Note that $\mathcal{E}_1$ and $\mathcal{E}_4$ promote respectively the arguments in favour of c_0 and those in favour of c_1, while $\mathcal{E}_2$ and $\mathcal{E}_3$ contain arguments supporting both classes.

We are now ready to define the new explanation function. For a given instance x, it returns the support of any argument in favour of $\mathtt{F}(x)$ that is in every stable extension and the support should be part of x. The intuition is the following: when two arguments cannot hold together (coherence being violated), both are discarded since at least one of them is incorrect. Without additional information (eg., on possible strength of arguments), it is hard to choose an argument at the expense of the other. The choice would certainly be arbitrary, and it is likely that the chosen argument is the wrong one. Our approach is therefore cautious as it abstains when facing a conflict between arguments. It keeps only the arguments that belong to all stable extensions.

Definition 20 (Explainer $\mathtt{L}^*$). *Let $\mathcal{Y} \subseteq \mathtt{Inst}$ and $x \in \mathcal{Y}$. The set of explanations of x is the following:*

$$\mathtt{L}^*(x) = \{H \subseteq \mathtt{Lit} \mid H \subseteq x \text{ and } \langle H, \mathtt{F}(x)\rangle \in \bigcap_{\mathcal{E}_i \in \sigma(AS)} \mathcal{E}_i\},$$

where $AS = \langle \arg(\mathcal{Y}), \mathcal{R}\rangle$ is an argumentation system.

Example 2 (Cont.) It can be checked that $\bigcap_{\mathcal{E}_i \in \sigma(AS)} \mathcal{E}_i = \emptyset$. Hence, $\forall x \in \mathcal{Y}$, $\mathtt{L}^*(x) = \emptyset$. This means that with the available information, it is not possible to generate reasonable abductive explanations.

In the above example, none of the instances of the dataset have explanations since all the five arguments are attacked. However, this is not always the case as shown in the following example.

Example 7. *Consider the theory below.*

$\mathcal{Y}$	f_1	f_2	$\mathtt{F}(I_i)$
x_1	*0*	*0*	c_1
x_2	*0*	*1*	c_2
x_3	*1*	*0*	c_3

The argumentation system that is built from $\mathcal{Y}$ has three arguments a_1, a_2, a_3:

- $a_1 = \langle\{L_1\}, c_1\rangle$ $\quad L_1 = \{(f_1, 0), (f_2, 0)\}$
- $a_2 = \langle\{L_2\}, c_2\rangle$ $\quad L_2 = \{(f_2, 1)\}$
- $a_3 = \langle\{L_3\}, c_3\rangle$ $\quad L_3 = \{(f_1, 1)\}$

Note that a_2 attacks a_3 and a_3 attacks a_2. Thus, the system has two stable extensions:

- $\mathcal{E}_1 = \{a_1, a_2\}$
- $\mathcal{E}_2 = \{a_1, a_3\}$

Thus, $\mathtt{L}^*(x_1) = \{L_1\}$ *and* $\mathtt{L}^*(x_2) = \mathtt{L}^*(x_3) = \emptyset$.

When $\mathtt{L}^*$ is applied on the whole feature space, the attack relation of the corresponding argumentation system would be empty, and the generated explanations coincide with the absolute ones. It is also clear that the function returns plausible explanations.

Proposition 20. *Let $\mathcal{Y} \subseteq \mathtt{Inst}$ and $x \in \mathcal{Y}$.*

- *If $\mathcal{Y} = \mathtt{Inst}$, then $\mathtt{L}^* = \mathtt{L}_{ae}$,*
- $\mathtt{L}^*(x) \subseteq \mathtt{L}_{pe}(x)$.

One can observe that the sole explanation of x_1 in Example 7 is the support of an argument (a_1) which is not attacked in the argumentation system. We characterize below the set of explanations returned by the novel function $\mathtt{L}^*$, and show that it only provides supports of non-attacked arguments. Before giving the result, let us first introduce some useful notation.

Notation: For $\mathcal{Y} \subseteq \mathtt{Inst}$, we denote by $\arg^*(\mathcal{Y})$ the set of all non-attacked arguments, i.e., $\arg^*(\mathcal{Y}) = \{a \in \arg(\mathcal{Y}) \mid \nexists b \in \arg(\mathcal{Y})$ such that b attacks $a\}$.

Theorem 5. *Let $\mathcal{Y} \subseteq \mathtt{Inst}$ and $x \in \mathcal{Y}$.*

$$\mathtt{L}^*(x) = \{H \subseteq \mathtt{Lit} \mid H \subseteq x \text{ and } \langle H, \mathtt{F}(x)\rangle \in \arg *(\mathcal{Y})\},$$

As a consequence, we show that $\mathtt{L}^*$ keeps among all plausible explanations (see Definition 15), those that are not involved in any conflict.

Corollary 1. *Let $\mathcal{Y} \subseteq \mathtt{Inst}$, $H \subseteq \mathtt{Lit}$ and $x \in \mathcal{Y}$. $H \in \mathtt{L}^*(x)$ iff:*

- *$H \in \mathtt{L}_{pe}(x)$, and*
- *$\forall x' \in \mathcal{Y}$ such that $\mathtt{F}(x) \neq \mathtt{F}(x')$, $\nexists H' \in \mathtt{L}_{pe}(x')$ with $H \cup H'$ is consistent.*

Let us now switch to the behaviour of the function $\mathtt{L}^*$ regarding the two properties of coherence and success. We show that it satisfies the former but violates the latter.

Proposition 21. *The function* $\mathtt{L}^*$ *satisfies coherence and violates success.*

To sum up, the function $\mathtt{L}^*$ violates success due mainly to the impossibility result given in Theorem 4. The function itself is very cautious as it considers only non-conflicting plausible explanations, rejecting thus all conflicts and leaving them unsolved. To explain more instances, a function would have to solve in a reasoned way conflicts, and for that it would need additional information. In the next section, we propose another function which may explain more instances than $\mathtt{L}^*$ while ensuring coherence. It solves conflicts using external information, namely priorities on features. Like $\mathtt{L}^*$, it uses intersection for aggregating extensions. Intersection ensures coherence by preventing picking up an argument from one extension and its attacker from another extension.

5.3. Weighted Explanation Function

In the previous section, we introduced a function which generates abductive explanations from a dataset while satisfying coherence at the cost of success. When facing a conflict, the function rejects all the involved arguments. In order to solve conflicts, additional information should be used to effectively discriminate between arguments. For example, one might assign to every argument a *strength.* The latter may represent different things including the proportion of instances of a dataset that are covered by the argument's support. The greater the proportion, the stronger the argument. In what follows, we rather investigate another source of strength, which comes from priorities on features. It is quite common in classification tasks that some features are more important for the outcomes than others. Hence, explanations referring to important features are more reliable than those referring to less important ones. It is worth mentioning that whatever the source of strength, the approach followed for defining an explainer taking them into account is similar to the one we describe below.

In what follows, we propose an explainer, denoted $\mathtt{wL}^*$, which assumes that each feature and each subset of features in $\mathcal{F} = \{f_1, \ldots, f_n\}$ has an *importance degree.* Such degrees are ascribed by a *capacity*, called also *fuzzy measure* in [44], which is a function that assigns to every subset of features a value from the unit interval $[0, 1]$.

Definition 21 (Capacity). *A* capacity *over a set* X *is a function* $\mathcal{V}$ *from* $\mathcal{P}(X)$[1] *to* $[0, 1]$ *satisfying the following conditions:*

[1] $\mathcal{P}(X)$ denotes the power set of the set X.

- $\mathcal{V}(\emptyset) = 0$ *(Boundary condition)*
- $\mathcal{V}(A) \leq \mathcal{V}(B)$ *whenever* $A \subseteq B \subseteq \texttt{Inst}$ *(Monotonicity)*

We call weighted theory a classical theory (Definition 2) whose set of features is equipped with a capacity.

Definition 22 (Weighted Theory). *A* weighted classification theory *is a pair* $(\texttt{T}, \mathcal{V})$ *where* $\texttt{T} = \langle \mathcal{F}, \texttt{dom}, \mathcal{C} \rangle$ *is a theory and* $\mathcal{V}$ *is a capacity over* $\mathcal{F}$*.*

We assume a set $\mathcal{Y} \subseteq \texttt{Inst}$ of instances and a classifier $\texttt{F}$ which can be queried on any instance in $\mathcal{Y}$. The novel explainer $\texttt{wL}^*$ starts by generating arguments from $\mathcal{Y}$ following Definition 14. Then, it assigns a strength to every argument in $\arg(\mathcal{Y})$. The strength is the importance degree of the features involved in the argument's support.

Definition 23 (Argument Strength). *Let* $a = \langle H, c \rangle \in \arg(\mathcal{Y})$*. The strength of* a *is* $\texttt{St}(a) = \mathcal{V}(\{f_i \mid (f_i, v_i) \in H\})$*.*

This strength is used for defining a defeat relation between arguments. It is based on the attack relation (Definition 17) but prevents an argument from attacking a stronger one, which makes defeat not symmetrical.

Definition 24 (Defeat Relation). *Let* $a, b \in \arg(\mathcal{Y})$*. We say that* a defeats b *iff:*

- *a attacks b, and*
- $\texttt{St}(a) \geq \texttt{St}(b)$*.*

We introduce now the extended version of an argumentation system.

Definition 25 (Weighted AS). *A* weighted argumentation system *built from* $\mathcal{Y} \subseteq \texttt{Inst}$ *is a pair* $AS = \langle \arg(\mathcal{Y}), \texttt{Def} \rangle$ *where* $\texttt{Def} \subseteq \arg(\mathcal{Y}) \times \arg(\mathcal{Y})$ *such that for* $a, b \in \arg(\mathcal{Y})$*,* $(a, b) \in \texttt{Def}$ *iff* a *defeats* b *(in the sense of Definition 24).*

Arguments are evaluated using stable semantics. We show that any weighted argumentation system has stable extensions since it does not contain elementary odd-length cycles. Furthermore, its extensions are a subset of those of the non-weighted system (Definition 18).

Proposition 22. *Let* $\mathcal{Y} \subseteq \texttt{Inst}$*,* $AS = \langle \arg(Y), \mathcal{R} \rangle$ *and* $AS' = \langle \arg(Y), \texttt{Def} \rangle$*.*

- $\sigma(AS') \neq \emptyset$,
- $\sigma(AS') \subseteq \sigma(AS)$.

For every instance, the new explainer $\mathtt{wL}^*$ returns the supports of arguments which belong to all stable extensions and are part of the instance.

Definition 26 (Explainer $\mathtt{wL}^*$). *Let $(\mathtt{T}, \mathcal{V})$ be a weighted theory, $\mathcal{Y} \subseteq \mathtt{Inst}$ and $x \in \mathcal{Y}$. The set of explanations of x is the following:*

$$\mathtt{wL}^*(x) = \{H \subseteq \mathtt{Lit} \mid H \subseteq x \text{ and } \langle H, \mathtt{F}(x)\rangle \in \bigcap_{\mathcal{E}_i \in \sigma(AS)} \mathcal{E}_i\},$$

where $AS = \langle \arg(\mathcal{Y}), \mathtt{Def}\rangle$ is a weighted argumentation system.

Example 7 (Cont.) Recall that there are three arguments that may be built from $\mathcal{Y}$ (a_1, a_2, a_3) with:

- $a_1 = \langle\{L_1\}, c_1\rangle$ $\qquad L_1 = \{(f_1, 0), (f_2, 0)\}$
- $a_2 = \langle\{L_2\}, c_2\rangle$ $\qquad L_2 = \{(f_2, 1)\}$
- $a_3 = \langle\{L_3\}, c_3\rangle$ $\qquad L_3 = \{(f_1, 1)\}$

Assume that the feature f_2 is more important than the feature f_1, i.e., $\mathcal{V}(\{f_1, f_2\}) \geq \mathcal{V}(\{f_2\}) > \mathcal{V}(\{f_1\})$. Hence, $\mathtt{St}(a_1) \geq \mathtt{St}(a_2) > \mathtt{St}(a_3)$. Thus, a_2 defeats a_3 while the converse does not hold. The corresponding weighted argumentation system has a single stable extension: $\mathcal{E} = \{a_1, a_2\}$. Consequently, $\mathtt{wL}^*(x_1) = \{L_1\}$, $\mathtt{wL}^*(x_2) = \{L_2\}$ and $\mathtt{wL}^*(x_3) = \emptyset$.

Note that the importance of features has been used for solving the conflict between the two arguments a_2 and a_3. Hence, unlike the function $\mathtt{L}^*$, the weighted $\mathtt{wL}^*$ selects a_2 and thus provides an explanation for the instance x_2. This function enriches thus the outcomes of $\mathtt{L}^*$ as confirmed by the following result.

Proposition 23. *Let $(\mathtt{T}, \mathcal{V})$ be a weighted theory, $\mathcal{Y} \subseteq \mathtt{Inst}$ and $x \in \mathcal{Y}$. It holds that: $\mathtt{L}^*(x) \subseteq \mathtt{wL}^*(x) \subseteq \mathtt{L}_{pe}(x)$. The converses do not hold.*

The above example shows that the weighted function $\mathtt{wL}^*$ violates success ($\mathtt{wL}^*(x_3) = \emptyset$). However, we show next that it satisfies coherence.

Proposition 24. *The function $\mathtt{wL}^*$ satisfies coherence and violates success.*

6. Related Work

As automated decision-making systems are becoming popular and deployed in several domains, the question of their explainability becomes increasingly important, leading to a growing literature related to explanation. Some of them focus on symbolic AI systems including reasoning models like argumentation (eg., [45]) and answer set programming (eg., [46]), recommendation systems (eg., [47]), multiple criteria decision systems (eg., [48]), and planning systems (eg., [49, 50]). Explanations in these works shed light on the system's internal decision process. Other works are interested in explaining machine learning models. Most of them are experimental, focusing on specific models, exposing their internal representations to find correlations *post hoc* between these representations and the predictions, and they are thus more about the arguably vaguer notion of interpretability. Furthermore, they focused more on local explanations (eg. [51]). Examples of such explanation functions are LIME [8], Anchors [9], SHAP [52] and EXPLAN [53]. They are mainly validated in an experimental way, and no formal guarantees are provided. In [54, 55], the authors tried to generate global explanations that provide insights in a black-box model's decision making process. For that purpose, they start by generating local explanations, then aggregate them using various operators. The approaches have been validated experimentally and the results have shown that explanations of LIME do not reliably represent model's global behaviour.

In our paper we investigated formal approaches for explainability that ensure formal guarantees. In the past few years, there is increasing interest in such approaches, and all existing works generate explanations from the whole feature space, which in practice is not reasonable ([12, 14, 18, 11, 16, 56]). They also all, with the exception of [14], investigated local explanations. They focused on abductive explanations and/or counterfactuals which they generate either for an arbitrary classifier (eg., [12, 14, 11]) or for specific ones like Bayesian networks, decision trees and random forests (eg., [18, 16]).

In our paper, we are more interested in explaining complex classifiers whose internal reasoning is difficult to grasp. Hence, we defined explanation functions which look for correlations between instances and the outcomes provided by classifiers. We provided a unified setting for representing different types of explanations including abductive and counterfactuals. We provided formal analysis of the links between global and local explanations. Furthermore, unlike the above-cited works, we also investigated formally functions that generate explanations from a subset of the feature space. We have shown that this raises particular challenges and explanation functions should be defined with great care in this context.

Unlike our work which explains existing black-box classifiers, [31, 57] proposed novel classification models that are based on arguments. Their explanations are defined in dialectical way as fictitious dialogues between a proponent (supporting an output) and an opponent (attacking the output) following [33]. The authors in [58, 59, 60] followed the same approach for defining explainable multiple decision systems, recommendation systems, or scheduling systems. In the above papers an argument is simply an instance and its label while our arguments pro/con are much richer. This shows that they are proposed for different purposes.

There is also a great interest in explaining the outcomes of argumentation frameworks (eg., [61, 62, 45]). The objective is to explain why an argument is accepted or, alternatively, rejected under a given semantics from [33]. These works are thus not related to ours as they do not focus on classifiers. Furthermore, since argumentation is interpretable, their explanations provide insight into the semantics while we consider classifiers as black-boxes.

7. Conclusion

This paper investigated the different notions of (local, global) explanation that have been discussed in the literature for interpreting black-box classifiers without "opening" them. It proposes the first formal setting for *defining, generating*, and *comparing* the most prominent types, namely *abductive*, *counterfactuals*, and *contrastive* explanations. The setting is based on two dual types of arguments (pros and cons) for justifying predictions. It used them as building blocks of explanations. This work lays the foundations for formal comparisons with other types of explanation.

In this paper, we also argued that generating explanations from the whole feature space is not reasonable in practice, and one should only consider a subset which may be chosen in different ways. However, generating explanations under such incomplete information raises particular issues, namely the possibility of incorrectness of explanations or non-existence of explanations. Thus, the definition of explainers should be done with great care. The paper provided the first two functions that ensure correctness while generating explanations from datasets. The functions are based on a well-known non-monotonic reasoning approach, namely argumentation.

This work can be extended in different ways. First, we have seen that the novel functions $\mathtt{L}^*$ and $\mathtt{wL}^*$ may return an empty set of explanations for an instance. While cautious reasoning is suitable when dealing with conflicting information by non-monotonic reasoning models, it may be a great weakness in XAI since a user would always expect an explanation for the outcome provided by a classifier. Hence, a future work consists of exploring other functions that would improve the outputs of $\mathtt{L}^*$ and $\mathtt{wL}^*$. Another line of research consists of using gradual semantics from [63] for evaluating arguments. Such semantics provide finer-grained evaluations.

Acknowledgements

Support from the ANR-3IA Artificial and Natural Intelligence Toulouse Institute is gratefully acknowledged.

8. Appendix: Proofs

Proof of Property 1 The three properties follow straightforwardly from the definition of an instance as a tuple of the n available attributes of $\mathcal{F}$. ■

Proof of Property 2 The property follows straightforwardly from the definition of $x_{\downarrow H}$. ■

Proof of Property 3 Let $\mathtt{G}$ and $\mathtt{L}$ be a class explainer and instance explainer respectively.

Assume that $\mathtt{G}$ and $\mathtt{L}$ are compatible and $\mathtt{G}$ is coherent. Let $x, x' \in \mathtt{Inst}$ such that $\mathtt{F}(x) \neq \mathtt{F}(x')$. From compatibility, $\mathtt{L}(x) \subseteq \mathtt{G}(\mathtt{F}(x))$, so $\forall H \in \mathtt{L}(x)$, $H \in \mathtt{G}(\mathtt{F}(x))$. Similarly, $\mathtt{L}(x') \subseteq \mathtt{G}(\mathtt{F}(x'))$, so $\forall H' \in \mathtt{L}(x')$, $H' \in \mathtt{G}(\mathtt{F}(x'))$. Since $\mathtt{F}(x) \neq \mathtt{F}(x')$, then from coherence of $\mathtt{G}$, $H \cup H'$ is inconsistent. Then, $\mathtt{L}$ is coherent.

Assume that $\mathtt{G}$ and $\mathtt{L}$ are compatible and $\mathtt{L}$ is coherent. Let $c, c' \in \mathcal{C}$ such that $c \neq c'$. From compatibility of $\mathtt{L}$ and $\mathtt{G}$, we have $\mathtt{G}(c) = \bigcup\limits_{x \in \mathtt{Inst} \text{ s.t. } \mathtt{F}(x)=c} \mathtt{L}(x)$ and $\mathtt{G}(c') = \bigcup\limits_{x \in \mathtt{Inst} \text{ s.t. } \mathtt{F}(x)=c'} \mathtt{L}(x)$. Let $H \in \mathtt{G}(c)$ and $H' \in \mathtt{G}(c')$. Then, $\exists x, x' \in \mathtt{Inst}$ such that $H \in \mathtt{L}(x)$ and $H' \in \mathtt{L}(x')$, $\mathtt{F}(x) = c$ and $\mathtt{F}(x') = c'$. From coherence of $\mathtt{L}$, $H \cup H'$ is inconsistent. ■

Proof of Property 4 Example 2 provides a counter example for $\mathtt{L}_{pe}(x) \subseteq \mathtt{L}_{ae}(x)$. Indeed, $U_3 = \{(C, 1),\ (E, 0)\}$ is a plausible explanation of x_5 in $\mathcal{Y} = \{x_1, \ldots, x_7\}$ while it is not plausible in $\mathcal{Y} \cup \{x_5\}$. ■

Proof of Property 5 The first property is straightforward from Definition 17. Let a be argument and c its conclusion. Assume that a attacks itself. From Definition 17, $c \neq c$ which is impossible. ■

Proof of Proposition 1 Let $\mathtt{T} = \langle \mathcal{F}, \mathtt{dom}, \mathcal{C} \rangle$ be a theory and $c \in \mathcal{C}$.

Let us show the equivalence. Assume that $\langle \emptyset, c \rangle \in \mathtt{arg}^+(\mathtt{T})$ for some $c \in \mathcal{C}$. Since $\forall x \in \mathtt{Inst}, \emptyset \subset x$, then by Definition 9, $\forall x \in \mathtt{Inst}, \mathtt{F}(x) = c$. Assume now that $\forall x \in \mathtt{Inst}, \mathtt{F}(x) = c$ (assumption (1)). We show that i) $\langle \emptyset, c \rangle \in \mathtt{arg}^+(\mathtt{T})$, and ii) if $\langle H', c' \rangle \in \mathtt{arg}^+(\mathtt{T})$, then $\langle H', c' \rangle = \langle \emptyset, c \rangle$. Since $\emptyset \subseteq \mathtt{Lit}$, $\emptyset$ is consistent, and $\forall x \in \mathtt{Inst}, \emptyset \subset x$, then $\langle \emptyset, c \rangle \in \mathtt{arg}^+(\mathtt{T})$. Assume now that $\langle H', c' \rangle \in \mathtt{arg}^+(\mathtt{T})$. Hence, H' is consistent. From Property 1, $\exists x \in \mathtt{Inst}$ such that $H' \subseteq x$. It follows that $\mathtt{F}(x) = c'$. From assumption (1), $c = c'$. Since $\langle \emptyset, c \rangle \in \mathtt{arg}^+(\mathtt{T})$, then $H' = \emptyset$ (otherwise minimality would be violated).

Let us show the second property. $T = \{x \in \mathtt{Inst} \mid \mathtt{F}(x) = c\}$ and $x \in T$. Since x is consistent (from Property 1), then $\forall H \subseteq x$, H is consistent too (from Property 1). So, $\exists H \subseteq x$ such that H is minimal (for set inclusion) such that $\forall x' \in \mathtt{Inst}$ s.t. $H \subseteq x'$, $\mathtt{F}(x') = c$. Note that $H = \emptyset$ iff $T = \mathtt{Inst}$, $H = x$ if $|T| = 1$, and $H \subseteq x$ otherwise. Hence, $\langle H, c \rangle \in \mathtt{arg}^+(\mathtt{T})$.

Let us show the third property. Assume that $\langle H, c \rangle \in \mathtt{arg}^+(\mathtt{T})$. From Definition 9, H is consistent. From Property 1, $\exists x \in \mathtt{Inst}$ s.t $H \subseteq x$. By Definition 9, $\mathtt{F}(x) = c$.

The last property follows straightforwardly from the two previous ones. ■

Proof of Proposition 2 Assume a classifier $\mathtt{F}$ is a surjective function, thus for every class $c \in \mathcal{C}$, there exists at least one $x \in \mathtt{Inst}$ such that $\mathtt{F}(x) = c$. From property 4 of Proposition 1, $\mathtt{Pros}(c) \neq \emptyset$. Thus, $\mathtt{G}_{pro}(c) \neq \emptyset$. ■

Proof of Proposition 3 Let $\mathtt{T} = \langle \mathcal{F}, \mathtt{dom}, \mathcal{C} \rangle$ be a theory and $c, c' \in \mathcal{C}$ with $c \neq c'$. Let also $\langle H, c \rangle, \langle H', c' \rangle \in \mathtt{arg}^+(\mathtt{T})$. Assume that $H \cup H'$ is consistent. Hence, $\exists x \in \mathtt{Inst}$ such that $H \cup H' \subseteq x$ (from item 2b of Property 1). Hence, $\mathtt{F}(x) = c$ and $\mathtt{F}(x) = c'$, which contradicts the fact that M is a function that ascribes a single class to every instance.

Let $H \in \mathtt{G}_{pro}(c)$ and $H' \in \mathtt{G}_{pro}(c')$. By definition of $\mathtt{G}_{pro}$, $\langle H, c \rangle \in \mathtt{Pros}(c)$ and $\langle H', c' \rangle \in \mathtt{Pros}(c')$. From the previous property, $H \cup H'$ is inconsistent. Thus, $\mathtt{G}_{pro}$ satisfies coherence. ■

Proof of Proposition 4 Let $\mathtt{F}$ be a classification model and $\mathtt{T} = \langle \mathcal{F}, \mathtt{dom}, \mathcal{C} \rangle$ a theory with $\mathcal{C} = \{c_1, \ldots, c_m\}$. For every $i \in \{1, \ldots, m\}$, let $\mathtt{Inst}_i = \{x \in \mathtt{Inst} \mid \exists \langle H, c_i \rangle \in \mathtt{arg}^+(\mathtt{T})$ and $H \subseteq x\}$.

Let us first show that $\mathtt{Inst}_i \cap \mathtt{Inst}_j = \emptyset$ for $i \neq j$. Assume that $x \in \mathtt{Inst}_i \cap \mathtt{Inst}_j$. By definition of $\mathtt{Inst}_i, \mathtt{Inst}_j$, there exist $\langle H, c_i \rangle, \langle H', c_j \rangle \in \mathtt{arg}^+(\mathtt{T})$ such that $H \subseteq x$ and $H' \subseteq x$. Thus, $H \cup H' \subseteq x$. But from Proposition 3, the set $H \cup H'$ is inconsistent while any instance $x \in \mathtt{Inst}$ is consistent (from Property 1).

Let us now show that $\texttt{Inst} = \texttt{Inst}_1 \cup \ldots \cup \texttt{Inst}_m$. Obviously, $\texttt{Inst}_1 \cup \ldots \cup \texttt{Inst}_m \subseteq \texttt{Inst}$. Let $x \in \texttt{Inst}$ and let us show that $x \in \texttt{Inst}_1 \cup \ldots \cup \texttt{Inst}_m$. From Definition 3, $\texttt{F}$ assigns a class from $\mathcal{C}$ to every instance in $\texttt{Inst}$. Hence, $\exists c_i \in \mathcal{C}$ such that $\texttt{F}(x) = c_i$. From Proposition 1, there exists $\langle H, c_i \rangle \in \texttt{arg}^+(\texttt{T})$. Hence, $x \in \texttt{Inst}_i$. ■

Proof of Proposition 5 In Example 3, the two classes c_1 and c_2 have the same reason for avoiding them, namely $\{(f_1, 1)\}$. The latter is consistent. ■

Proof of Proposition 6 Let $\texttt{T} = \langle \mathcal{F}, \texttt{dom}, \mathcal{C} \rangle$ be a theory such that $\mathcal{C} = \{c, c'\}$. Let $\langle H, c \rangle \in \texttt{Pros}(c)$. From Definition 9, $H \subseteq \texttt{Lit}$ is consistent and minimal (for set inclusion) such that: $\forall x \in \texttt{Inst}$, if $H \subseteq x$, then $\texttt{F}(x) = c$, thus $\texttt{F}(x) \neq c'$. By Definition 10, it follows that $\langle H, \overline{c'} \rangle \in \texttt{Cons}(c')$. Following the same reasoning, we show that if $\langle H, \overline{c'} \rangle \in \texttt{Cons}(c')$, then $\langle H, c \rangle \in \texttt{Pros}(c)$. ■

Proof of Proposition 7 Let $\texttt{T} = \langle \mathcal{F}, \texttt{dom}, \mathcal{C} \rangle$ be a theory and $c \in \mathcal{C}$. Let $\langle H, c \rangle \in \texttt{arg}^+(\texttt{T})$ and $\langle H', \overline{c} \rangle \in \texttt{arg}^-(\texttt{T})$ such that $H \cup H'$ is consistent. From Property 1, $\exists x \in \texttt{Inst}$ such that $H \cup H' \subseteq x$. From Definition 9, $\texttt{F}(x) = c$ and from Definition 10 $\texttt{F}(x) \neq c$. Since $\texttt{F}$ assigns a class to ever instance, then $\exists c' \in \mathcal{C}$ such that $c \neq c'$ and $\texttt{F}(x) = c'$. This contradicts the fact that $\texttt{F}$ is a function that assigns a single class to every instance. ■

Proof of Proposition 8 Let $\texttt{T} = \langle \mathcal{F}, \texttt{dom}, \mathcal{C} \rangle$ be a theory and $c \in \mathcal{C}$.

Let us show the equivalence. Assume that $\langle \emptyset, \overline{c} \rangle \in \texttt{Cons}(c)$. Thus, $\forall x \in \texttt{Inst}$, $\emptyset \subset x$, $\texttt{F}(x) \neq c$. Assume now that $\forall x \in \texttt{Inst}$, $\texttt{F}(x) \neq c$. Clearly, $\emptyset \subset \texttt{Lit}$, $\emptyset$ is consistent and minimal (for set inclusion) such that $\emptyset \subset x$, for any $x \in \texttt{Inst}$. Hence, by Definition 10 it follows that $\langle \emptyset, \overline{c} \rangle \in \texttt{Cons}(c)$.

Let us show the second property. Let $\langle H, \overline{c} \rangle \in \texttt{arg}^-(\texttt{T})$. By Definition 10, H is consistent. From Property 1, $\exists x \in \texttt{Inst}$ such that $H \subseteq x$. By Definition 10, $\texttt{F}(x) \neq c$.

Let us show the third property. Assume that $\exists x^* \in \texttt{Inst}$ s.t. $\texttt{F}(x^*) \neq c$. Since $\texttt{F} : \texttt{Inst} \to \mathcal{C}$, then $\exists c' \in \mathcal{C}$ s.t. $c \neq c'$ and $\texttt{F}(x^*) = c'$. From Proposition 1, $\exists \langle H, c' \rangle \in \texttt{arg}^+(\texttt{T})$ and $H \subseteq x^*$. Note that for any $x \in \texttt{Inst}$ s.t. $H \subseteq x$, $\texttt{F}(x) = c'$, thus $\texttt{F}(x) \neq c$. Let $T = \{x \in \texttt{Inst} \mid \texttt{F}(x) = c' \text{ and } H \subseteq x\}$. If $\forall x \in \texttt{Inst} \setminus T$, $\texttt{F}(x) = c$, then $\langle H, \overline{c} \rangle \in \texttt{arg}^-(\texttt{T})$ since any strict subset of H would not be sufficient for getting c', thus for avoiding c. Otherwise, $\exists H' \subseteq H$ that is sufficient for avoiding c and $\langle H', \overline{c} \rangle \in \texttt{arg}^-(\texttt{T})$.

Let us show the fourth property. Let $c \neq c'$ and $\langle H', c' \rangle \in \texttt{arg}^+(\texttt{T})$. There are two cases: i) $H' = \emptyset$. From Proposition 1, for all $x \in \texttt{Inst}$,

$\mathtt{F}(x) = c'$, hence $\mathtt{F}(x) \neq c$. Thus, $\langle \emptyset, \overline{c} \rangle \in \mathtt{Cons}(c)$ (from above equivalence). ii) $H' \neq \emptyset$. By Definition 9, $\forall x \in \mathtt{Inst}$ s.t. $H' \subseteq x$, $\mathtt{F}(x) = c'$, i.e., $\mathtt{F}(x) \neq c$. Since H' is consistent, then $\forall H \subset H'$, H is consistent. Hence, let $H \subseteq H'$ be the minimal (for set inclusion) subset such that for every $x \in X$ s.t. $H \subseteq x$, $\mathtt{F}(x) \neq c$. Hence, $\langle H, \overline{c} \rangle \in \mathtt{arg}^{-}(\mathtt{T})$ with $\emptyset \subseteq H \subseteq H'$. ■

Proof of Proposition 9 Let $\mathtt{T} = \langle \mathcal{F},\ \mathtt{dom},\ \mathcal{C} \rangle$ be a theory and $c \in \mathcal{C}$. Let us show the first equivalence. Assume that $\mathtt{Pros}(c) = \emptyset$. Hence, $\nexists \langle H, c \rangle \in \mathtt{arg}^{+}(\mathtt{T})$. From Proposition 1, $\nexists x \in \mathtt{Inst}$ s.t. $\mathtt{F}(x) = c$, thus $\forall x \in \mathtt{Inst}$, $\mathtt{F}(x) \neq c$. From Proposition 8, $\langle \emptyset, \overline{c} \rangle \in \mathtt{Cons}(c)$. Assume that $\exists \langle H, \overline{c} \rangle \in \mathtt{Cons}(c)$. Obviously, $H = \emptyset$ because $\langle \emptyset, \overline{c} \rangle \in \mathtt{Cons}(c)$, otherwise, H would violate the minimality condition. Hence, $\mathtt{Cons}(c) = \{\langle \emptyset, \overline{c} \rangle\}$.

Assume now that $\mathtt{Cons}(c) = \{\langle \emptyset, \overline{c} \rangle\}$. From the first item of Proposition 8, $\forall x \in \mathtt{Inst}$, $\mathtt{F}(x) \neq c$. From the fourth item of Proposition 1, $\mathtt{Pros}(c) = \emptyset$.

We show the second equivalence. Assume that $\mathtt{Cons}(c) = \emptyset$. Thus, $\nexists \langle H, \overline{c} \rangle \in \mathtt{arg}^{-}(\mathtt{T})$. From the third item of Proposition 8, $\nexists x \in \mathtt{Inst}$ s.t. $\mathtt{F}(x) \neq c$, i.e., $\forall x \in \mathtt{Inst}$, it holds that $\mathtt{F}(x) = c$. From the first item of Proposition 1, $\mathtt{arg}^{+}(\mathtt{T}) = \{\langle \emptyset, c \rangle\} = \mathtt{Pros}(c)$.

Assume now that $\mathtt{Pros}(c) = \{\langle \emptyset, c \rangle\}$. From the first item of Proposition 1, $\forall x \in \mathtt{Inst}$, $\mathtt{F}(x) = c$, i.e., $\nexists x \in \mathtt{Inst}$ s.t. $\mathtt{F}(x) \neq c$. From the second item of Proposition 8, $\nexists \langle H, \overline{c} \rangle \in \mathtt{arg}^{-}(\mathtt{T})$, thus $\mathtt{Cons}(c) = \emptyset$. ■

Proof of Proposition 10 Let $c \in \mathcal{C}$. Let us show that $\mathcal{Y} \subseteq \mathtt{Inst} \setminus \mathcal{Z}$. Let $x \in \mathcal{Y}$. By definition of $\mathcal{Y}$, $\exists \langle H, \overline{c} \rangle \in \mathtt{arg}^{-}(\mathtt{T})$ and $H \subseteq x$. Assume that $x \in \mathcal{Z}$, hence $\exists \langle H', c \rangle \in \mathtt{arg}^{+}(\mathtt{T})$ and $H' \subseteq x$. So, $H \cup H' \subseteq x$. From Proposition 7, $H \cup H'$ is inconsistent while x is consistent from Property 1. Consequently, $x \notin \mathcal{Z}$ and so $x \in \mathtt{Inst} \setminus \mathcal{Z}$.

Assume now that $x \in \mathtt{Inst} \setminus \mathcal{Z}$, so $x \notin \mathcal{Z}$. From Proposition 4, $\mathtt{F}(x) \neq c$. From Proposition 8, $\exists \langle H, \overline{c} \rangle \in \mathtt{arg}^{-}(\mathtt{T})$ and $H \subseteq x$. Thus, $x \in \mathcal{Y}$. ■

Proof of Proposition 11 Let $c \in \mathcal{C}$. From Proposition 9, $(\mathtt{Cons}(c) = \emptyset) \iff (\mathtt{Pros}(c) = \{\langle \emptyset, c \rangle\})$. From item 1 of Proposition 1, $\mathtt{Pros}(c) = \{\langle \emptyset, c \rangle\}$ iff $\forall x \in \mathtt{Inst}$, $\mathtt{F}(x) = c$. But $|\mathcal{C}| > 1$ and $\mathtt{F}$ is surjective. Thus, $\exists y \in \mathtt{Inst}$ such that $\mathtt{F}(y) \neq c$. Then, $\mathtt{Pros}(c) \neq \{\langle \emptyset, c \rangle\}$ and so $\mathtt{Cons}(c) \neq \emptyset$. Consequently, $\mathtt{G}_{con}(c) \neq \emptyset$. ■

Proof of Proposition 12 Let $\mathtt{T} = \langle \mathcal{F},\ \mathtt{dom},\ \mathcal{C} \rangle$ be a theory, $x \in \mathtt{Inst}$ and $c \in \mathcal{C}$ such that $\mathtt{F}(x) = c$.

Assume that $\mathtt{L}_{ae}(x) = \{\emptyset\}$. Then, $\emptyset \in \mathtt{G}_{pro}(c)$ and so $\langle \emptyset, c \rangle \in \mathtt{Pros}(c)$. From item 1 of Proposition 1, $\forall y \in \mathtt{Inst}$, $\mathtt{F}(y) = c$ (1). Assume now that (1) holds. From item 1 of Proposition 1, $\langle\emptyset, c\rangle \in \texttt{Pros}(c)$ and thus $\emptyset \in \texttt{G}_{pro}(c)$. Hence, $\emptyset \in \texttt{L}_{ae}(x)$. Suppose that $\exists H \in \texttt{L}_{ae}(x)$, so $\langle H, c\rangle \in \texttt{Pros}(c)$. It follows that $H = \emptyset$ since $\langle\emptyset, c\rangle \in \texttt{Pros}(c)$, otherwise H would violate minimality.

The second property is straightforward from the definition. ■

Proof of Proposition 13 Let $\texttt{T} = \langle\mathcal{F}, \texttt{dom}, \mathcal{C}\rangle$ be a theory, $x \in \texttt{Inst}$ and $c \in \mathcal{C}$ s.t. $\texttt{F}(x) = c$. From item 2 of Proposition 1, $\exists\langle H, c\rangle \in \texttt{Pros}(c)$ and $H \subseteq x$. Hence, $H \in \texttt{G}_{pro}(c)$ and from Definition 11, $H \in \texttt{L}_{ae}(x)$ and so $\texttt{L}_{ae}(x) \neq \emptyset$ meaning that $\texttt{L}_{ae}$ satisfies success.

Let us now show that $\texttt{L}_{ae}$ is coherent. Let $x, x' \in \texttt{Inst}$ such that $\texttt{F}(x) \neq \texttt{F}(x')$. Assume that $H \in \texttt{L}_{ae}(x)$ and $H' \in \texttt{L}_{ae}(x')$. From Definition 11, $H \in \texttt{G}_{pro}(\texttt{F}(x))$ and $H' \in \texttt{G}_{pro}(\texttt{F}(x'))$. From Proposition 3, it follows that $H \cup H'$ is inconsistent.

Let us show that $\texttt{L}_{ae}$ and $\texttt{G}_{pro}$ are compatible. From Proposition 12, we have for any $x \in \texttt{Inst}$, $\texttt{L}_{ae}(x) \subseteq \texttt{G}_{pro}(c)$. Let us now show that $\texttt{G}_{pro}(c) \subseteq \bigcup\limits_{x \in \texttt{Inst} \text{ s.t. } \texttt{F}(x)=c} \texttt{L}_{ae}(x)$. Let $H \in \texttt{G}_{pro}(c)$. By Definition 9, $\exists\langle H, c\rangle \in \texttt{Pros}(c)$. From item 3 of Proposition 1, $\exists x \in \texttt{Inst}$ such that $\texttt{F}(x) = c$. Thus, from Definition 11, $H \in \texttt{L}_{ae}(x)$. ■

Proof of Proposition 14 Let $x \in \texttt{Inst}$ such that $\texttt{F}(x) = c$.

The first implication follows straightforwardly from Definition 12.

Assume that $\texttt{L}_{cf}(x) = \emptyset$. So, $\texttt{Cons}(c) = \emptyset$. From Propositions 1 and 9, $\forall y \in \texttt{Inst}$, $\texttt{F}(y) = c$. Assume now that $\forall y \in \texttt{Inst}$, $\texttt{F}(y) = c$. From Proposition 1, $\texttt{Pros}(c) = \{\langle\emptyset, c\rangle\}$. From Proposition 9, $\texttt{Cons}(c) = \emptyset$. Hence, $\texttt{L}_{cf}(x, c) = \emptyset$.

Assume that $\emptyset \in \texttt{L}_{cf}(x, c)$. Thus, $\langle\emptyset, \overline{c}\rangle \in \texttt{Cons}(c)$. From Proposition 9, $\texttt{Pros}(c) = \emptyset$. From Proposition 1, $\forall y \in \texttt{Inst}$, $\texttt{F}(y) \neq c$. This contradicts the fact $\texttt{F}(x) = c$. ■

Proof of Proposition 15 Example 3 shows that the instance explainer $\texttt{L}_{cf}$ violates coherence. Note that $\texttt{F}(x_1) = c_1$ and $\texttt{F}(x_2) = c_2$ while the set $\{(f_1, 1)\}$ is a common counterfactual to both instances.

If the classifier $\texttt{F}$ is surjective, then $\forall c \in \mathcal{C}$, $\exists x \in \texttt{Inst}$ such that $\texttt{F}(x) = c$. Hence, from Proposition 14 $\texttt{L}_{cf}(x) \neq \emptyset$. ■

Proof of Proposition 16 Let $\texttt{T} = \langle\mathcal{F}, \texttt{dom}, \mathcal{C}\rangle$ be a theory, $\mathcal{C} = \{c, c'\}$, and $x \in \texttt{Inst}$ such that $\texttt{F}(x) = c$. From Theorem 3, $H \in \texttt{L}_{cf}(x)$ iff $H \subseteq \texttt{Lit}$ is subset-minimal such that $\texttt{F}(x_{\downarrow H}) \neq \texttt{F}(x)$. Since $\mathcal{C} = \{c, c'\}$, it follows that $\texttt{F}(x_{\downarrow H}) = c'$. Hence, $H \in \texttt{L}_{co}(x, c')$. ■

Proof of Proposition 17 Example 3 shows that the instance explainer $\mathtt{L}_{co}$ violates coherence. Note that $\mathtt{F}(x_1) = c_1$ and $\mathtt{F}(x_2) = c_2$ while the set $\{(f_1, 1)\}$ is a common contrastive to (x_1, c_3) and (x_2, c_3). If the classifier $\mathtt{F}$ is surjective, then $\forall c \in \mathcal{C}$, $\exists x \in \mathtt{Inst}$ such that $\mathtt{F}(x) = c$. Hence, $\mathtt{L}_{co}(x) \neq \emptyset$. ■

Proof of Proposition 18 A plausible explanation is a part of an instance. Every instance is consistent, then its subparts are all consistent. The second property is straightforward. The last property is straightforward. ■

Proof of Proposition 19 Let $x \in \mathcal{Y}$. Since x is consistent, then $\exists H \subseteq x$ such that H is minimal (for set inclusion) such that $\forall y \in \mathcal{Y}$, if $H \subseteq y$, then $\mathtt{F}(y) = \mathtt{F}(x)$. Hence, $\mathtt{L}_{pe}(x) \neq \emptyset$.

In order to show that $\mathtt{L}_{pe}$ violates coherence, let us consider again Example 2. Recall that there are two classes in the theory: c_0, c_1. Their arguments are given below:

- $a_1 = \langle U_1, c_0 \rangle \qquad U_1 = \{(V, 0)\}$
- $a_2 = \langle U_2, c_0 \rangle \qquad U_2 = \{(M, 1)\}$
- $a_3 = \langle U_3, c_0 \rangle \qquad U_3 = \{(C, 1), (E, 0)\}$
- $a_4 = \langle U_4, c_1 \rangle \qquad U_4 = \{(V, 1)\}$
- $a_5 = \langle U_5, c_1 \rangle \qquad U_5 = \{(M, 0)\}$

It can be checked that:

- $\mathtt{L}_{pe}(x_1) = \{U_1, U_2\}$
- $\mathtt{L}_{pe}(x_5) = \{U_1, U_2, U_3\}$
- $\mathtt{L}_{pe}(x_2) = \mathtt{L}_{pe}(x_4) = \mathtt{L}_{pe}(x_7) = \{U_4, U_5\}$

Note that $U_1 \cup U_5$ is consistent while $\mathtt{F}(x_1) \neq \mathtt{F}(x_2)$ is consistent. ■

Proof of Proposition 20 Assume that $\mathcal{Y} = \mathtt{Inst}$ and $x \in \mathtt{Inst}$. The inclusion $\mathtt{L}^*(x) \subseteq \mathtt{L}_{pe}(x)$ follows from Definition 20.

By Definition 14, $\arg(\mathcal{Y}) = \arg^+(\mathtt{T})$. Furthermore, from Proposition 3, for all $c, c' \in \mathcal{C}$ with $c \neq c'$, for all $\langle H, c \rangle, \langle H', c' \rangle \in \arg^+(\mathtt{T})$, the set $H \cup H'$ is inconsistent. Hence, the attack relation is empty and thus there exists a single stable extension $\mathcal{E} = \arg^+(\mathtt{T})$. Then, $\mathtt{L}^*(x) = \mathtt{L}_{ae}(x)$. ■

Proof of Proposition 21 Example 2 shows that success is violated. Indeed, $\mathtt{L}^*(x_1) = \emptyset$. Assume now that $\mathtt{L}^*$ violates Coherence. Thus, there exist $x, x' \in \mathcal{Y}$, there exist $H \in \mathtt{L}^*(x)$ and $H' \in \mathtt{L}^*(x')$ such that $\mathtt{F}(x) \neq \mathtt{F}(x')$ and $H \cup H'$ is consistent. Note that $a = \langle H, \mathtt{F}(x)\rangle, b = \langle H', \mathtt{F}(x')\rangle \in \arg(\mathcal{Y})$ (by definition of $\mathtt{L}^*$). Furthermore, a and b attack each other, and $a, b \in \bigcap\limits_{\mathcal{E} \in \sigma(AS)} \mathcal{E}$, where $AS = \langle \arg(\mathcal{Y}), \mathcal{R}\rangle$. This contradicts the fact every stable extension is conflict-free. ■

Proof of Proposition 22 Let $\mathcal{Y} \subseteq \mathtt{Inst}$, $AS = \langle \arg(Y), \mathcal{R}\rangle$ and $AS' = \langle \arg(Y), \mathtt{Def}\rangle$.

Assume that AS' contains an elementary odd-length cycle, which is a sequence $\mathcal{A} = \{a_1, \ldots, a_{2k+1}\} \subseteq \arg(\mathcal{Y})$ such that:

i) $\forall i \in \{1, \ldots, 2k\}$, a_i defeats a_{i+1},

ii) a_{2k+1} defeats a_1,

iii) $\forall i \in \{1, \ldots, 2k+1\}$, $|\{x \in \mathcal{A} \mid x \text{ defeats } a_i\}| = 1$.

From i), a_1 attacks a_2, a_2 attacks $a_3, \ldots,$ a_{2k} attacks a_{2k+1}. From iii), $\mathtt{St}(a_1) > \mathtt{St}(a_2) > \ldots, \mathtt{St}(a_{2k+1})$. Thus, $a_1 > a_{2k+1}$, which contradicts ii) (i.e., the fact that a_{2k+1} defeats a_1 while $\mathtt{St}(a_{2k+1}) > \mathtt{St}(a_1)$ due to iii)). Thus, AS does not contain odd-length cycles. From [33], it follows that AS has a non-empty set of stable extensions.

Let us now show the inclusion $\sigma(AS') \subseteq \sigma(AS)$. Let $\mathcal{E} \in \sigma(AS')$. By definition of stable extensions, $\mathcal{E}$ is conflict-free and $\forall a \in \arg(\mathcal{Y}) \setminus \mathcal{E}$, $\exists b \in \mathcal{E}$ such that b defeats a. Thus, b attacks a. Let us show that $\mathcal{E}$ is conflict-free wrt $\mathcal{R}$. Assume $a, b \in \mathcal{E}$ such that a attacks b. Thus, either $\mathtt{St}(a) > \mathtt{St}(b)$ and so a defeats b, or $\mathtt{St}(b) > \mathtt{St}(a)$ and b defeats a, or $\mathtt{St}(a) = \mathtt{St}(b)$. The three cases contradicts the conflict-freeness of $\mathcal{E}$ in AS'. Then, $\mathcal{E} \in \sigma(AS)$. ■

Proof of Proposition 23 Let $\mathcal{Y} \subseteq \mathtt{Inst}$ and $x \in \mathcal{Y}$.

Assume $H \in \mathtt{wL}^*(x)$, then by definition of $\mathtt{wL}^*$, $H \subseteq x$ and $\langle H, \mathtt{F}(x)\rangle \in \arg(\mathcal{Y})$. So, $H \in \mathtt{L}_{pe}(x)$, which shows the inclusion $\mathtt{wL}^*(x) \subseteq \mathtt{L}_{pe}(x)$.

Assume now $H \in \mathtt{L}^*(x)$. Then, $H \subseteq x$ and $\langle H, \mathtt{F}(x)\rangle \in \bigcap\limits_{\mathcal{E}_i \in \sigma(AS)} \mathcal{E}_i$, where $AS = \langle \arg(\mathcal{Y}), \mathcal{R}\rangle$. Let $AS' = \langle \arg(\mathcal{Y}), \mathtt{Def}\rangle$. From the second property in Proposition 22, $\sigma(AS') \subseteq \sigma(AS)$, so $\bigcap\limits_{\mathcal{E}_i \in \sigma(AS)} \mathcal{E}_i \subseteq \bigcap\limits_{\mathcal{E}_j \in \sigma(AS')} \mathcal{E}_j$. It follows that $\langle H, \mathtt{F}(x)\rangle \in \bigcap\limits_{\mathcal{E}_j \in \sigma(AS')} \mathcal{E}_j$ and $H \in \mathtt{wL}^*(x)$, which shows the inclusion $\mathtt{L}^*(x) \subseteq \mathtt{wL}^*(x)$. ■

Proof of Proposition 24 Example 7 shows that success is violated as $\mathtt{L}^*(x_3) = \emptyset$. Assume now that $\mathtt{wL}^*$ violates Coherence. Thus, there exist $x, x' \in \mathcal{Y}$ such that $\mathtt{F}(x) \neq \mathtt{F}(x')$ and there exist $H \in \mathtt{wL}^*(x)$ and $H' \in \mathtt{wL}^*(x')$ such that $H \cup H'$ is consistent. Let $a = \langle H, \mathtt{F}(x)\rangle$ and $b = \langle H', \mathtt{F}(x')\rangle$. Note that $a, b \in \arg(\mathcal{Y})$ (by definition of $\mathtt{wL}^*$) and a attacks b and b attacks a. There are three cases: i) $\mathtt{St}(a) > \mathtt{St}(b)$, then a defeats b, ii) $\mathtt{St}(b) > \mathtt{St}(a)$, then b defeats a, or iii) $\mathtt{St}(a) = \mathtt{St}(b)$, then a defeats b and b defeats a. By Definition 26. $\{a, b\} \subseteq \bigcap_{\mathcal{E}_i \in \sigma(AS)} \mathcal{E}_i$, where $AS = \langle \arg(\mathcal{Y}), \mathtt{Def}\rangle$. This contradicts the fact that every stable extension is conflict-free. ■

Proof of Theorem 1 Let $\mathtt{T} = \langle \mathcal{F},\ \mathtt{dom},\ \mathcal{C}\rangle$ be a theory and $c \in \mathcal{C}$. Let $\mathtt{Supp}(c) = \{H_1, \ldots, H_k\}$ be such that for every $H_i \in \mathtt{Supp}(c)$, the following hold:

- $H_i \subseteq \mathtt{Lit}$
- H_i is consistent
- $\forall H \in \mathtt{G}_{con}(c)$, $H_i \cup H$ is inconsistent,
- $\nexists H' \subseteq \mathtt{Lit}$ such that $H' \subset H_i$ and H' satisfies the third condition.

Let us first show that $\mathtt{Supp}(c) \subseteq \mathtt{G}_{pro}(c)$. Let $H \in \mathtt{Supp}(c)$, we show that $\langle H, c\rangle \in \mathtt{Pros}(c)$ (hence, $H \in \mathtt{G}_{pro}(c)$). From definition of $\mathtt{Supp}(c)$, $H \subseteq \mathtt{Lit}$ and H is consistent.
We show now that for any $x \in \mathtt{Inst}$ such that $H \subseteq x$, $\mathtt{F}(x) = c$. Assume that for some $x \in \mathtt{Inst}$, $H \subseteq x$ and $\mathtt{F}(x) \neq c$. From item 3 of Proposition 8, $\exists\langle H', \overline{c}\rangle \in \mathtt{arg}^-(\mathtt{T})$ such that $H' \subseteq x$. Thus, $H \cup H' \subseteq x$. By definition of $\mathtt{Supp}(c)$, $H \cup H'$ is inconsistent, which contradicts the fact that x is consistent (from item 1 of Property 1).
Let us show the minimality condition of an argument. Assume that $\exists H' \subset H$ such that $\forall x \in \mathtt{Inst}$, if $H' \subseteq x$ then $\mathtt{F}(x) = c$. Due to the minimality condition in the definition of $\mathtt{Supp}(c)$, $\exists\langle H'', \overline{c}\rangle \in \mathtt{arg}^-(\mathtt{T})$ such that $H'' \cup H$ is consistent. From item 2b in Property 1, $\exists x \in \mathtt{Inst}$ such that $H'' \cup H \subseteq x$. Hence, $\mathtt{F}(x) = c$ (due to H') and $\mathtt{F}(x) \neq c$ (due to $\langle H'', \overline{c}\rangle$). This contradicts the fact that $\mathtt{F}$ assigns a single class to every instance.

Let us now show that $\mathtt{G}_{pro}(c) \subseteq \mathtt{Supp}(c)$. Let $\langle H, c\rangle \in \mathtt{Pros}(c)$ and we show that $H \in \mathtt{Supp}(c)$. There are two cases: i) $H = \emptyset$ and ii) $H \neq \emptyset$. If $H = \emptyset$, then from Proposition 9, $\mathtt{Cons}(c) = \emptyset$. Obviously, $\emptyset \subseteq \mathtt{Lit}$ and $\emptyset$ is consistent. Furthermore, the third condition of the definition of $\mathtt{Supp}(c)$ is satisfied in a vacuous way, and $\emptyset$ is the minimal set that satisfies it. Hence,

$\emptyset \in \texttt{Supp}(c)$. If $H \neq \emptyset$, then from Proposition 9, $\texttt{Cons}(c) \neq \emptyset$. From Proposition 7, for any $\langle H_i, \overline{c}\rangle \in \texttt{Cons}(c)$, $H \cup H_i$ is inconsistent. Finally, assume that $\exists H' \subset H$ such that H' is the smallest subset such that for any $\langle H_i, \overline{c}\rangle \in \texttt{Cons}(c)$, $H' \cup H_i$ is inconsistent. Then, $H' \in \texttt{Supp}(c)$. From the above implication, $\langle H', c\rangle \in \texttt{Pros}(c)$. This contradicts the fact that $\langle H, c\rangle \in \texttt{Pros}(c)$ as H would violate minimality. ■

Proof of Theorem 2 Let $\texttt{T} = \langle \mathcal{F}, \texttt{dom}, \mathcal{C}\rangle$ be a theory and $c \in \mathcal{C}$. Let $\texttt{Att}(c) = \{H_1, \ldots, H_k\}$ be such that for every $i = 1, \ldots, k$,

- $H_i \subseteq \texttt{Lit}$
- H_i is consistent
- $\forall H \in \texttt{G}_{pro}(c)$, $H \cup H_i$ is inconsistent
- $\nexists H' \subseteq \texttt{Lit}$ such that $H' \subset H_i$ and H' satisfies the third condition.

We start by showing that $\texttt{Att}(c) \subseteq \texttt{G}_{con}(c)$. Let $H \in \texttt{Att}(c)$, and we show that $\langle H, \overline{c}\rangle \in \texttt{Cons}(c)$ (thus $H \in \texttt{G}_{con}(c)$). From the definition of $\texttt{Att}(c)$, $H \subseteq \texttt{Lit}$ and H is consistent. We show now that for any $x \in \texttt{Inst}$ such that $H \subseteq x$, $\texttt{F}(x) \neq c$. Assume that for some $x \in \texttt{Inst}$, $H \subseteq x$ and $\texttt{F}(x) = c$. From item 2 of Proposition 1, $\exists \langle H', c\rangle \in \texttt{Pros}(c)$ and $H' \subseteq x$. Thus, $H \cup H' \subseteq x$. By the definition of $\texttt{Att}(c)$, $H \cup H'$ is inconsistent, which contradicts the fact that x is consistent. Let us show the minimality condition of an argument. Assume that $\exists H' \subset H$ such that H' is the smallest subset such that $\forall x \in \texttt{Inst}$, if $H' \subseteq x$, $\texttt{F}(x) \neq c$. From definition of $\texttt{Att}(c)$, $\exists \langle H'', c\rangle \in \texttt{Pros}(c)$ such that $H'' \cup H'$ is consistent. From item 2b of Property 1, $\exists x \in \texttt{Inst}$ such that $H'' \cup H' \subseteq x$. Hence, $\texttt{F}(x) = c$ (due to $\langle H'', c\rangle \in \texttt{Pros}(c)$) and $\texttt{F}(x) \neq c$ (due to H').

We now show that $\texttt{G}_{con}(c) \subseteq \texttt{Att}(c)$. We consider $\langle H, \overline{c}\rangle \in \texttt{Cons}(c)$, and show that $H \in \texttt{Att}(c)$. If $H = \emptyset$, then from Proposition 9, $\texttt{Pros}(c) = \emptyset$. Obviously, $\emptyset \subseteq \texttt{Lit}$ and $\emptyset$ is consistent. Furthermore, the third condition of the definition of $\texttt{Att}(c)$ is satisfied in a vacuous way, and $\emptyset$ is the minimal set that satisfies it. Hence, $\emptyset \in \texttt{Att}(c)$. If $H \neq \emptyset$, then from Proposition 9, $\texttt{Pros}(c) \neq \emptyset$. From Proposition 7, for any $\langle H_i, c\rangle \in \texttt{Pros}(c)$, $H \cup H_i$ is inconsistent. Finally, assume that $\exists H' \subset H$ such that H' is the smallest subset such that for any $\langle H_i, c\rangle \in \texttt{Pros}(c)$, $H' \cup H_i$ is inconsistent. Then, $H' \in \texttt{Att}(c)$. From the above implication, $\langle H', \overline{c}\rangle \in \texttt{Cons}(c)$. This contradicts the fact that $\langle H, \overline{c}\rangle \in \texttt{Cons}(c)$ as H would violate minimality. ■

Proof of Theorem 3 Let $x \in \texttt{Inst}$ such that $\texttt{F}(x) = c$.

Let $H \in \mathtt{L}_{cf}(x)$, thus $\exists U \in \mathtt{Cons}(c)$ s.t. $H = U \setminus x$ (a) and $\nexists U' \in \mathtt{Cons}(c)$ such that $U' \setminus x \subset H$ (b). Since by definition U is consistent, then from Property 1 H is consistent. From Property 2, $y = x_{\downarrow H} \in \mathtt{Inst}$. Furthermore, $\mathtt{F}(y) \neq c$ since $U \subseteq y$. Let us now assume some $H' \subset H$ such that $\mathtt{F}(x_{\downarrow H'}) \neq c$. Let $z = x_{\downarrow H'}$. From Proposition 8, $\exists U' \in \mathtt{Cons}(c)$ such that $H' \subseteq z$. Let $U'_1 = U' \cap H'$ and $U'_2 = U' \setminus H'$. Since $U'_2 \subseteq x$, then $H'_1 \neq \emptyset$. Hence, $U' \setminus x = U'_1 \subset H' \subset U \setminus x$, which contradicts the assumption (b).

Let $H \subseteq \mathtt{Lit}$ be a minimal for set inclusion such that H is consistent and $\mathtt{F}(x_{\downarrow H}) \neq \mathtt{F}(x)$. Let $y = x_{\downarrow H}$. From Proposition 8, $\exists U \in \mathtt{Cons}(\mathtt{F}(x))$ such that $U \subseteq y$. Since $\mathtt{F}(x) = c$, then $H \cap U \neq \emptyset$. Let $U = U_1 \cup U_2$ such that $U_1 = H \cap U$ and $U_2 = U \setminus H$. Assume that $H \neq U_1$ (i.e., $U_1 \subset H$). From Property 1, U_1 is consistent (being a subset of a consistent set U), then $\exists z \in \mathtt{Inst}$ such that $z = x_{\downarrow U_1}$. Note that $U_2 \subseteq z$ since $U_2 \subseteq x$, hence $U \subseteq z$ and so $\mathtt{F}(z) \neq c$. This contradicts the minimality of H. ■

Proof of Theorem 4 Recall that Coherence and Success are *compatible* iff there exists a plausible explainer, say Ł, which satisfies both properties. Recall also that Ł satisfies Coherence (resp. Success) iff the property holds for every theory, every dataset and every classifier. To show that Coherence and Success are *not compatible*, it is sufficient to show that such a function Ł does not exist.

Assume that $\mathtt{L}$ is a plausible explainer that satisfies both Coherence and Success. Consider the theory below made of two binary features f_1, f_2, and a binary classifier $\mathtt{F}$. The table below summarizes the predictions made by the classifier for the simple dataset $\mathcal{Y}$.

$\mathcal{Y}$	f_1	f_2	$\mathtt{F}(I_i)$
x_1	0	1	0
x_2	1	0	1

The function $\mathtt{L}_{pe}$ returns the following plausible explanations.

- $\mathtt{L}_{pe}(x_1) = \{L_1\}$ $\qquad L_1 = \{(f_2, 1)\}$
- $\mathtt{L}_{pe}(x_2) = \{L_2\}$ $\qquad L_2 = \{(f_1, 1)\}$

Since $\mathtt{L}$ is a plausible explainer, then from Definition 16 it holds that:

$$\forall i \in \{1, 2\}, \quad \mathtt{L}(x_i) \subseteq \mathtt{L}_{pe}(x_i) \qquad \text{(A1)}.$$

Since $\mathtt{L}$ satisfies Success, then $\mathtt{L}(x_1) \neq \emptyset$ and $\mathtt{L}(x_2) \neq \emptyset$. Thus, $\forall i \in \{1, 2\}$, $\mathtt{L}(x_i) = \mathtt{L}_{pe}(x_i)$. However, $L_1 \cup L_2$ is consistent while $\mathtt{F}(x_1) \neq \mathtt{F}(x_2)$, thus $\mathtt{L}$ violates Coherence.

Let us now start by coherence. From coherence of $\mathtt{L}$, $\nexists L, L' \in \mathtt{L}(x_1) \cup \mathtt{L}(x_2)$ such that $L \cup L'$ is consistent, $L \in \mathtt{L}(x_i)$, $L' \in \mathtt{L}(x_j)$, and $\mathtt{F}(x_i) \neq \mathtt{F}(x_j)$ (A2). From (A1), $\mathtt{L}(x_1) \cup \mathtt{L}(x_2) \subseteq \mathtt{L}_{pe}(x_1) \cup \mathtt{L}_{pe}(x_2)$. But, $\mathtt{L}_{pe}(x_1) \cup \mathtt{L}_{pe}(x_2) = \{L_1, L_2\}$, then from (A2) either $L_1 \notin \mathtt{L}(x_1) \cup \mathtt{L}(x_2))$, in which case $\mathtt{L}(x_1) = \emptyset$, or $L_3 \notin \mathtt{L}(x_1) \cup \mathtt{L}(x_2)$, in which case $\mathtt{L}(x_2) = \emptyset$. Thus, $\mathtt{L}$ violates Success. ■

Proof of Theorem 5 Let $\mathcal{Y} \subseteq \mathtt{Inst}$ and $x \in \mathcal{Y}$. Let $AS = \langle \arg(\mathcal{Y}), \mathcal{R} \rangle$ be the argumentation system built from $\mathcal{Y}$. From Property 5, the attack relation $\mathcal{R}$ is symmetric and irreflexive. So from [43], $\forall a \in \arg(Y)$,

$$a \in \bigcap_{\mathcal{E} \in \sigma(AS)} \mathcal{E} \text{ iff } \{b \in \arg(\mathcal{Y}) \mid (b, a) \in \mathcal{R}\} = \emptyset.$$

Recall that $\{b \in \arg(\mathcal{Y}) \mid (b, a) \in \mathcal{R}\} = \arg^*(\mathcal{Y})$. So, from Definition 20, $\mathtt{L}^*(x) = \{H \mid \exists \langle H, \mathtt{F}(x) \rangle \in \arg^*(\mathcal{Y}) \text{ and } H \subseteq x\}$. ■

Proof of Corollary 1 Let $\mathcal{Y} \subseteq \mathtt{Inst}$, $H \subseteq \mathtt{Lit}$ and $x \in \mathcal{Y}$.

Assume that $H \in \mathtt{L}^*(x)$. Then, from Theorem 5 $\langle H, \mathtt{F}(x) \rangle \in \arg^*(\mathcal{Y})$ (**A1**). From Proposition 20, $H \in \mathtt{L}_{pe}(x)$. Let $x' \in \mathcal{Y}$ such that $\mathtt{F}(x) \neq \mathtt{F}(x')$. Assume that $\exists H' \in \mathtt{L}_{pe}(x')$ such that $H \cup H'$ is consistent. From Definition 15, $\langle H', \mathtt{F}(x') \rangle \in \arg(\mathcal{Y})$. Clearly, $\langle H', \mathtt{F}(x') \rangle$ attacks $\langle H, \mathtt{F}(x) \rangle$, which contradicts (**A1**).

Assume now that H satisfies the following conditions:

i) $H \in \mathtt{L}_{pe}(x)$, and

ii) $\forall x' \in \mathcal{Y}$ such that $\mathtt{F}(x) \neq \mathtt{F}(x')$, $\nexists H' \in \mathtt{L}_{pe}(x')$ with $H \cup H'$ is consistent.

From Definition 15 of $\mathtt{L}_{pe}$, $H \subseteq x$ and $\langle H, \mathtt{F}(x) \rangle \in \arg(\mathcal{Y})$.

Assume that $\langle H, \mathtt{F}(x) \rangle \notin \arg^*(\mathcal{Y})$. Thus, $\exists \langle H', c \rangle \in \arg(\mathcal{Y})$ such that $\langle H, \mathtt{F}(x) \rangle$ attacks $\langle H', c \rangle$, i.e., $H \cup H'$ is consistent and $\mathtt{F}(x) \neq c$. By Definition 14, $\exists y \in \mathcal{Y}$ such that $H' \subseteq y$ and $\mathtt{F}(y) = c$. Thus, $H' \in \mathtt{L}_{pe}(y)$, which contradicts the condition ii). Hence, $\langle H, \mathtt{F}(x) \rangle \in \arg^*(\mathcal{Y})$ and from Theorem 5, $H \in \mathtt{L}^*(x)$. ■

References

[1] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI Workshop on Explainable Artificial Intelligence (XAI), 2017, pp. 1–6.

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (5) (2019) 93:1–93:42.

[3] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[4] C. Molnar, Interpretable Machine Learning, Lulu.com, 2020.
URL `https://books.google.fr/books?id=RHjTxgEACAAJ`

[5] N. Burkart, M. Huber, A survey on the explainability of supervised machine learning, Journal of Artificial Intelligence Research 70 (2021) 245–317.

[6] I. Stepin, J. M. Alonso, A. Catalá, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.

[7] K. Cyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI, 2021, pp. 4392–4399.

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

[9] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 1527–1535.

[10] A. Dhurandhar, P. Chen, R. Luss, C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the missing: Towards contrastive explanations with pertinent negatives, in: Annual Conference on Neural Information Processing Systems, NeurIPS, 2018, pp. 590–601.

[11] A. Darwiche, A. Hirth, On the reasons behind decisions, in: 24th European Conference on Artificial Intelligence, ECAI, Vol. 325 of Frontiers in Artificial Intelligence and Applications, IOS Press, 2020, pp. 712–720.

[12] A. Ignatiev, N. Narodytska, J. Marques-Silva, Abduction-based explanations for machine learning models, in: The Thirty-Third Conference on Artificial Intelligence, AAAI, 2019, pp. 1511–1519.

[13] L. Amgoud, Explaining black-box classification models with arguments, in: 33rd IEEE International Conference on Tools with Artificial Intelligence, ICTAI, 2021, pp. 791–795.

[14] A. Ignatiev, N. Narodytska, J. Marques-Silva, On relating explanations and adversarial examples, in: Thirty-third Conference on Neural Information Processing Systems, NeurIPS, 2019, pp. 15857–15867.

[15] A. Ignatiev, N. Narodytska, N. Asher, J. Marques-Silva, From contrastive to abductive explanations and back again, in: AIxIA 2020 - Advances in Artificial Intelligence - XIXth International Conference of the Italian Association for Artificial Intelligence, Vol. 12414 of Lecture Notes in Computer Science, Springer, 2020, pp. 335–355.

[16] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, P. Marquis, On preferred abductive explanations for decision trees and random forests, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI, 2022, p. In press.

[17] L. Amgoud, Non-monotonic explanation functions, in: Proceedings of the 16th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU, Vol. 12897 of Lecture Notes in Computer Science, 2021, pp. 19–31.

[18] A. Shih, A. Choi, A. Darwiche, A symbolic approach to explaining Bayesian network classifiers, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI, 2018, pp. 5103–5111.

[19] A. Ignatiev, J. P. M. Silva, SAT-based rigorous explanations for decision lists, in: 24th International Conference on Theory and Applications of Satisfiability Testing - SAT, 2021, pp. 251–269.

[20] J. Ferreira, M. de Sousa Ribeiro, R. Gonçalves, J. Leite, Looking inside the black-box: Logic-based explanations for neural networks, in: 19th International Conference on Principles of Knowledge Representation and Reasoning, KR, 2022, p. In press.

[21] O. Biran, K. R. McKeown, Human-centric justification of machine learning predictions, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI, 2017, pp. 1461–1467.

[22] R. Luss, P. Chen, A. Dhurandhar, P. Sattigeri, K. Shanmugam, C. Tu, Generating contrastive explanations with monotonic attribute functions, CoRR (2019).
URL `http://arxiv.org/abs/1905.12698`

[23] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in AI, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 279–288.

[24] S. Wachter, B. D. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, CoRR abs/1711.00399 (2017).

[25] R. Byrne, Semifactual "even if" thinking, Thinking and reasoning 8 (1) (2002) 41–67.

[26] I. Rahwan, G. Simari(eds.), Argumentation in Artificial Intelligence, Springer, 2009.

[27] F. Lin, Y. Shoham, Argument systems - an uniform basis for non-monotonic reasoning, in: Proc. of KR, 1989, pp. 245 – 255.

[28] G. Simari, R. Loui, A mathematical treatment of defeasible reasoning and its implementation, Artificial Intelligence 53 (2-3) (1992) 125–157.

[29] P. Besnard, A. Hunter, A logic-based theory of deductive arguments, Artificial Intelligence 128 (1-2) (2001) 203–235.

[30] L. Amgoud, H. Prade, Using arguments for making and explaining decisions, Artificial Intelligence 173 (3-4) (2009) 413–436.

[31] L. Amgoud, M. Serrurier, Agents that argue and explain classifications, Journal of Autonomous Agents and Multi-Agent Systems 16 (2) (2008) 187–209.

[32] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. R. Simari, M. Thimm, S. Villata, Towards artificial argumentation, AI Magazine 38 (3) (2017) 25–36.

[33] P. M. Dung, On the Acceptability of Arguments and its Fundamental Role in Non-Monotonic Reasoning, Logic Programming and n-Person Games, Artificial Intelligence 77 (1995) 321–357.

[34] S. Kotsiantis, D. Kanellopoulos, Discretization techniques: A recent survey, GESTS International Transactions on Computer Science and Engineering 32(1) (2006) 47–58.

[35] D. Park, L. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal explanations: Justifying decisions and pointing to the evidence, CoRR abs/1802.08129 (2018).

[36] A. Schulz, F. Hinder, B. Hammer, Deepview: Visualizing classification boundaries of deep neural networks as scatter plots using discriminative dimensionality reduction, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI, 2020, pp. 2305–2311.

[37] O. Li, H. Liu, C. Chen, C. Rudin, Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI), 2018, pp. 3530–3537.

[38] I. Stepin, A. Catala, M. Pereira-Fariña, J. Alonso, Paving the way towards counterfactual generation in argumentative conversational agents, in: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI), Association for Computational Linguistics, 2019, pp. 20–25.

[39] Y. Dimopoulos, S. Dzeroski, A. Kakas, Integrating explanatory and descriptive learning in ILP, in: Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI, 1997, pp. 900–907.

[40] A. Kakas, F. Riguzzi, Abductive concept learning, New Generation Computing 18 (3) (2000) 243–294.

[41] R. Byrne, Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI, 2019, pp. 6276–6282.

[42] R. Byrne, Counterfactual thought, Annual Review of Psychology 67 (2016).

[43] S. Coste-Marquis, C. Devred, P. Marquis, Symmetric argumentation frameworks, in: Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU, 2005, pp. 317–328.

[44] G. Choquet, Theory of capacities, Annales de l'Institut Fourier 5 (1953) 131–295.

[45] B. Liao, L. van der Torre, Explanation semantics for abstract argumentation, in: H. Prakken, S. Bistarelli, F. Santini, C. Taticchi (Eds.), Computational Models of Argument - Proceedings of COMMA, Vol. 326 of Frontiers in Artificial Intelligence and Applications, IOS Press, 2020, pp. 271–282.

[46] J. Fandinno, C. Schulz, Answering the "why" in answer set programming - A survey of explanation approaches, Theory and Practice of Logic Programming 19 (2) (2019) 114–203.

[47] A. Rago, O. Cocarascu, C. Bechlivanidis, D. A. Lagnado, F. Toni, Argumentative explanations for interactive recommendations, Artificial Intelligence 296 (2021) 103506.

[48] C. Labreuche, Explanation with the winter value: Efficient computation for hierarchical choquet integrals, Int. J. Approx. Reason. 151 (2022) 225–250.

[49] B. Krarup, S. Krivic, D. Magazzeni, D. Long, M. Cashmore, D. E. Smith, Contrastive explanations of plans through model restrictions, Journal of Artificial Intelligence Research 72 (2021) 533–612.

[50] D. Aineto, E. Onaindia, M. Ramírez, E. Scala, I. Serina, Explaining the behaviour of hybrid systems with PDDL+ planning, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI, ijcai.org, 2022, pp. 4567–4573.

[51] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, 2017, pp. 4765–4774.

[52] S. M. Lundberg, G. G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles, ArXiv abs/1802.03888 (2018).

[53] P. Rasouli, I. C. Yu, EXPLAN: explaining black-box classifiers using adaptive neighborhood generation, in: 2020 International Joint Conference on Neural Networks, IJCNN, IEEE, 2020, pp. 1–9.

[54] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, Glocalx - from local to global explanations of black box ai models, Artificial Intelligence 294 (2021) 103457.

[55] I. van der Linden, H. Haned, E. Kanoulas, Global aggregations of local explanations for black box models, CoRR abs/1907.03039 (2019).
URL `http://arxiv.org/abs/1907.03039`

[56] R. Boumazouza, F. C. Alili, B. Mazure, K. Tabia, ASTERYX: A model-agnostic sat-based approach for symbolic and score-based explanations, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), The 30th ACM International Conference on Information and Knowledge Management, CIKM, ACM, 2021, pp. 120–129.

[57] O. Cocarascu, A. Stylianou, K. Cyras, F. Toni, Data-empowered argumentation for dialectically explainable predictions, in: Proceedings of the Twenty-Fourth Conference on Artificial Intelligence, ECAI, 2020, p. In press.

[58] K. Cyras, D. Birch, Y. Guo, F. Toni, R. Dulay, S. Turvey, D. Greenberg, T. Hapuarachchi, Explanations by arbitrated argumentative dispute, Expert Systems with Applications 127 (2019) 141–156.

[59] K. Cyras, D. Letsios, R. Misener, F. Toni, Argumentation for explainable scheduling, in: The Thirty-Third Conference on Artificial Intelligence, AAAI, 2019, pp. 2752–2759.

[60] A. Rago, O. Cocarascu, F. Toni, Argumentation-based recommendations: Fantastic explanations and how to find them, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI, 2018, pp. 1949–1955.

[61] A. Borg, F. Bex, Contrastive explanations for argumentation-based conclusions, in: 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS, 2022, pp. 1551–1553.

[62] A. Borg, F. Bex, Necessary and sufficient explanations for argumentation-based conclusions, in: J. Vejnarová, N. Wilson (Eds.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 16th European Conference, ECSQARU, Vol. 12897 of Lecture Notes in Computer Science, Springer, 2021, pp. 45–58.

[63] L. Amgoud, D. Doder, S. Vesic, Evaluation of argument strength in attack graphs: Foundations and semantics, Artificial Intelligence 302 (2022) 103607.