ON RELATING 'WHY?' AND 'WHY NOT?' EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Explanations of Machine Learning (ML) models often address a 'Why?' question. Such explanations can be related with selecting feature-value pairs which are sufficient for the prediction. Recent work has investigated explanations that address a 'Why Not?' question, i.e. finding a change of feature values that guarantee a change of prediction. Given their goals, these two forms of explaining predictions of ML models appear to be mostly unrelated. However, this paper demonstrates otherwise, and establishes a rigorous formal relationship between 'Why?' and 'Why Not?' explanations. Concretely, the paper proves that, for any given instance, 'Why?' explanations are minimal hitting sets of 'Why Not?' explanations and vice-versa. Furthermore, the paper devises novel algorithms for extracting and enumerating both forms of explanations.

1 INTRODUCTION

The importance of devising mechanisms for computing explanations of Machine Learning (ML) models cannot be overstated, as illustrated by the fast-growing body of work in this area. A glimpse of the importance of explainable AI (XAI) is offered by a growing number of recent surveys and overviews Hoffman & Klein (2017); Hoffman et al. (2017); Biran & Cotton (2017); Montavon et al. (2018); Klein (2018); Hoffman et al. (2018a); Adadi & Berrada (2018); Alonso et al. (2018); Dosilovic et al. (2018); Hoffman et al. (2018b); Guidotti et al. (2019); Samek et al. (2019); Samek & Müller (2019); Miller (2019b;a); Anjomshoae et al. (2019); Mittelstadt et al. (2019); Xu et al. (2019).

Past work on computing explanations has mostly addressed *local* (or instance-dependent) explanations Ribeiro et al. (2016); Lundberg & Lee (2017); Ribeiro et al. (2018); Shih et al. (2018; 2019); Ignatiev et al. (2019a); Darwiche & Hirth (2020); Darwiche (2020). Exceptions include for example approaches that distill ML models, e.g. the case of NNs Frosst & Hinton (2017) among many others Ribeiro et al. (2016), or recent work on relating explanations with adversarial examples Ignatiev et al. (2019b), both of which can be seen as seeking *global* (or instance-independent) explanations. Prior research has also mostly considered model-agnostic explanations Ribeiro et al. (2016); Lundberg & Lee (2017); Ribeiro et al. (2018). Recent work on model-based explanations, e.g. Shih et al. (2018); Ignatiev et al. (2019a), refers to local (or global) model-agnostic explanations as *heuristic*, given that these approaches offer no *formal* guarantees with respect to the underlying ML model[1]. Examples of heuristic approaches include Ribeiro et al. (2016); Lundberg & Lee (2017); Ribeiro et al. (2018), among many others[2]. In contrast, local (or global) model-based explanations are referred to as *rigorous*, since these offer the strongest formal guarantees with respect to the underlying ML model. Concrete examples of such rigorous approaches include Shih et al. (2018); Tran & d'Avila Garcez (2018); Shih et al. (2019); Ignatiev et al. (2019a;b); Darwiche & Hirth (2020); Jha et al. (2019).

Most work on computing explanations aims to answer a 'Why prediction π?' question. Some work proposes approximating the ML model's behavior with a linear model Ribeiro et al. (2016); Lundberg & Lee (2017). Most other work seeks to find a (often minimal) set of feature value pairs which is sufficient for the prediction, i.e. as long as those features take the specified values, the prediction does not change. For rigorous approaches, the answer to a 'Why prediction π?' question has been referred to as PI-explanations Shih et al. (2018; 2019), abductive explanations Ignatiev et al. (2019a), but also as (minimal) sufficient reasons Darwiche & Hirth (2020); Darwiche (2020). (Hereinafter, we use the term *abductive explanation* because of the other forms of explanations studied in the paper.)

[1]A taxonomy of ML model explanations used in this paper is included in Appendix A.

[2]There is also a recent XAI service offered by Google: `https://cloud.google.com/explainable-ai/`, inspired on similar ideas Google (2019).

Another dimension of explanations, studied in recent work Miller (2019b), is the difference between explanations for 'Why prediction π?' questions, e.g., 'Why did I get the loan?', and for 'Why prediction π and not δ?' questions, e.g., 'Why didn't I get the loan?'. Explanations for 'Why Not?' questions, labelled by Miller (2019b) *contrastive* explanations, isolate a pragmatic component of explanations that *abductive explanations* lack. Concretely, an abductive explanation identifies a set of feature values which are sufficient for the model to make a prediction π and thus provides an answer to the question 'Why π?' A constrastive explanation sets up a counterfactual link between what was a (possibly) *desired* outcome of a certain set of features and what was the observed outcome Bromberger (1962); Achinstein (1980). Thus, a contrastive explanation answers a 'Why π and not δ?' question Miller (2018); Dhurandhar et al. (2018); Mittelstadt et al. (2019).

In this paper we focus on the relationship between *local* abductive and contrastive explanations[3]. One of our contributions is to show how recent approaches for computing rigorous abductive explanations Shih et al. (2018; 2019); Ignatiev et al. (2019a); Darwiche & Hirth (2020); Darwiche (2020) can also be exploited for computing contrastive explanations. To our knowledge, this is new. In addition, we demonstrate that rigorous (model-based) local abductive and contrastive explanations are related by a minimal hitting set relationship [4], which builds on the seminal work of Reiter in the 80s Reiter (1987). Crucially, this novel hitting set relationship reveals a wealth of algorithms for computing and for enumerating contrastive and abductive explanations. We emphasize that it allows designing the first algorithm to *enumerate* abductive explanations. Finally, we demonstrate feasibility of our approach experimentally. Furthermore, our experiments show that there is a strong correlation between contrastive explanations and explanations produced by the commonly used SHAP explainer.

2 PRELIMINARIES

Explainability in Machine Learning. The paper assumes an ML model $\mathbb{M}$, which is represented by a finite set of first-order logic (FOL) sentences $\mathcal{M}$. (When applicable, simpler alternative representations for $\mathcal{M}$ can be considered, e.g. (decidable) fragments of FOL, (mixed-)integer linear programming, constraint language(s), etc.)[5] A set of features $\mathcal{F} = \{f_1, \ldots, f_L\}$ is assumed. Each feature f_i is categorical (or ordinal), with values taken from some set D_i. An *instance* is an assignment of values to features. The space of instances, also referred to as *feature* (or *instance*) *space*, is defined by $\mathbb{F} = D_1 \times D_2 \times \ldots \times D_L$. (For real-valued features, a suitable interval discretization can be considered.) A (feature) literal λ_i is of the form $(f_i = v_i)$, with $v_i \in D_i$. In what follows, a literal will be viewed as an atom, i.e. it can take value *true* or *false*. As a result, an instance can be viewed as a set of L literals, denoting the L distinct features, i.e. an instance contains a single occurrence of a literal defined on any given feature. A set of literals is consistent if it contains at most one literal defined on each feature. A consistent set of literals can be interpreted as a conjunction or as a disjunction of literals; this will be clear from the context. When interpreted as a conjunction, the set of literals denotes a *cube* in instance space, where the unspecified features can take any possible value of their domain. When interpreted as a disjunction, the set of literals denotes a *clause* in instance space. As before, the unspecified features can take any possible value of their domain.

The remainder of the paper assumes a classification problem with a set of classes $\mathbb{K} = \{\kappa_1, \ldots, \kappa_M\}$. A prediction $\pi \in \mathbb{K}$ is associated with each instance $X \in \mathbb{F}$. Throughout this paper, an ML model $\mathbb{M}$ will be associated with some logical representation (or encoding), whose consistency depends on the (input) instance and (output) prediction. Thus, we define a predicate $\mathcal{M} \subseteq \mathbb{F} \times \mathbb{K}$, such that $\mathcal{M}(X, \pi)$ is true iff the input X is consistent with prediction π given the ML model $\mathbb{M}$[6]. We further simplify the notation by using $\mathcal{M}_\pi(X)$ to denote a predicate $\mathcal{M}(X, \pi)$ for a concrete prediction π.

Moreover, we will compute *prime implicants* of $\mathcal{M}_\pi$. These predicates defined on $\mathbb{F}$ and represented as consistent conjunctions (or alternatively as sets) of feature literals. Concretely, a consistent

[3]In contrast with recent work Ignatiev et al. (2019b), which studies the relationship between *global* model-based (abductive) explanations and adversarial examples.

[4]A local abductive (resp. contrastive) explanation is a minimal hitting set of the set of all local contrastive (resp. abductive) explanations.

[5]$\mathcal{M}$ is referred to as the (formal) model of the ML model $\mathbb{M}$. The use of FOL is not restrictive, with fragments of FOL being used in recent years for modeling ML models in different settings. These include NNs Ignatiev et al. (2019a) and Bayesian Network Classifiers Shih et al. (2019), among others.

[6]This alternative notation is used for simplicity and clarity with respect to earlier work Shih et al. (2018); Ignatiev et al. (2019a;b). Furthermore, defining $\mathcal{M}$ as a predicate allows for multiple predictions for the same point in feature space. Nevertheless, such cases are not considered in this paper.

conjunction of feature literals τ is an implicant of $\mathcal{M}_\pi$ if the following FOL statement is true:

$$\forall(X \in \mathbb{F}).\tau(X) \rightarrow \mathcal{M}(X, \pi) \tag{1}$$

The notation $\tau \vDash \mathcal{M}_\pi$ is used to denote that τ an implicant of $\mathcal{M}_\pi$. Similarly, a consistent set of feature literals ν is the negation of an implicate of $\mathcal{M}_\pi$ if the following FOL statement is true:

$$\forall(X \in \mathbb{F}).\nu(X) \rightarrow (\vee_{\rho \neq \pi} \mathcal{M}(X, \rho)) \tag{2}$$

$\mathcal{M}_\pi \vDash \neg\nu$, or alternatively $(\nu \vDash \neg\mathcal{M}_\pi) \equiv (\nu \vDash \vee_{\rho \neq \pi} \mathcal{M}_\rho)$. An implicant τ (resp. implicate ν) is called *prime* if none of its proper subsets $\tau' \subsetneq \tau$ (resp. $\nu' \subsetneq \nu$) is an implicant (resp. implicate).

Abductive explanations represent prime implicants of the decision function associated with some predicted class π[7].

Analysis of Inconsistent Formulas. Throughout the paper, we will be interested in formulas $\mathcal{F}$ that are *inconsistent* (or *unsatisfiable*), i.e. $\mathcal{F} \vDash \bot$, represented as conjunctions of clauses. Some clauses in $\mathcal{F}$ can be *relaxed* (i.e. allowed not to be satisfied) to restore consistency, whereas others cannot. Thus, we assume that $\mathcal{F}$ is partitioned into two first-order subformulas $\mathcal{F} = \mathcal{B} \cup \mathcal{R}$, where $\mathcal{R}$ contains the *relaxable* clauses, and $\mathcal{B}$ contains the *non-relaxable* clauses. $\mathcal{B}$ can be viewed as (consistent) background knowledge, which must always be satisfied.

Given an inconsistent formula $\mathcal{F}$, represented as a set of first-order clauses, we identify the clauses that are responsible for unsatisfiability among those that can be relaxed, as defined next[8].

Definition 1 (Minimal Unsatisfiable Subset (MUS)) *Let $\mathcal{F} = \mathcal{B} \cup \mathcal{R}$ denote an inconsistent set of clauses ($\mathcal{F} \vDash \bot$). $\mathcal{U} \subseteq \mathcal{R}$ is a* Minimal Unsatisfiable Subset *(MUS) iff $\mathcal{B} \cup \mathcal{U} \vDash \bot$ and $\forall_{\mathcal{U}' \subsetneq \mathcal{U}}$, $\mathcal{B} \cup \mathcal{U}' \nvDash \bot$.*

Informally, an MUS provides the minimal information that needs to be added to the background knowledge $\mathcal{B}$ to obtain an inconsistency; it explains the causes for this inconsistency. Alternatively, one might be interested in correcting the formula, removing some clauses in $\mathcal{R}$ to achieve consistency.

Definition 2 (Minimal Correction Subset (MCS)) *Let $\mathcal{F} = \mathcal{B} \cup \mathcal{R}$ denote an inconsistent set of clauses ($\mathcal{F} \vDash \bot$). $\mathcal{T} \subseteq \mathcal{R}$ is a* Minimal Correction Subset *(MCS) iff $\mathcal{B} \cup \mathcal{R} \setminus \mathcal{T} \nvDash \bot$ and $\forall_{\mathcal{T}' \subsetneq \mathcal{T}}$, $\mathcal{B} \cup \mathcal{R} \setminus \mathcal{T}' \vDash \bot$.*

A fundamental result in reasoning about inconsistent clause sets is the minimal hitting set (MHS) duality relationship between MUSes and MCSes Reiter (1987); Birnbaum & Lozinskii (2003): *MCSes are MHSes of MUSes and vice-versa.* This result has been extensively used in the development of algorithms for MUSes and MCSes Bailey & Stuckey (2005); Liffiton & Sakallah (2008); Liffiton et al. (2016), and also applied in a number of different settings. Recent years have witnessed the proposal of a large number of novel algorithms for the extraction and enumeration of MUSes and MCSes Bacchus & Katsirelos (2015); Liffiton et al. (2016); Grégoire et al. (2018); Bendík et al. (2018). Although most work addresses propositional theories, these algorithms can easily be generalized to any other setting where entailment is monotonic, e.g. SMT de Moura & Bjørner (2008).

Running Example. The following example will be used to illustrate the main ideas.

Example 1 *We consider a textbook example Poole & Mackworth (2010)[Figure 7.1, page 289] addressing the classification of a user's preferences regarding whether to read or to skip a given book. For this dataset, the set of features is:*

$$\{\ \mathsf{A(uthor)}, \mathsf{T(hread)}, \mathsf{L(ength)}, \mathsf{W(hereRead)}\ \}$$

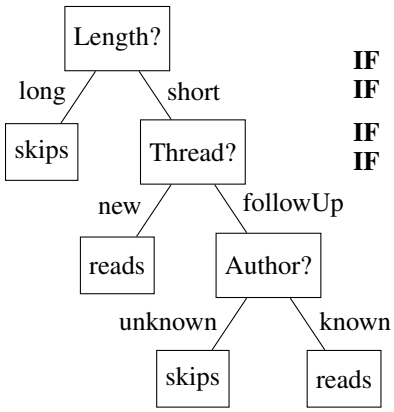
All features take one of two values, respectively $\{\mathsf{known}, \mathsf{unknown}\}$, $\{\mathsf{new}, \mathsf{followUp}\}$, $\{\mathsf{long}, \mathsf{short}\}$, *and* $\{\mathsf{home}, \mathsf{work}\}$. *An example instance is:* $\{(\mathsf{A} = \mathsf{known}), (\mathsf{T} = \mathsf{new}), (\mathsf{L} = \mathsf{long}), (\mathsf{W} = \mathsf{home})\}$ *This instance is identified as e_1 Poole & Mackworth (2010) with prediction* skips. *Figure 1a shows a possible decision tree for this example Poole & Mackworth (2010)*[9]. *The decision tree can be represented as a set of rules as shown in Figure 1b*[10].

[7]By definition of prime implicant, abductive explanations are sufficient reasons for the prediction. Hence the names used in recent work: abductive explanations Ignatiev et al. (2019a), PI-explanations Shih et al. (2018; 2019) and sufficient reasons Darwiche & Hirth (2020); Darwiche (2020).

[8]The definitions in this section are often presented for the propositional case, but the extension to the first-order case is straightforward.

[9]The choice of a decision tree aims only at keeping the example(s) presented in the paper as simple as possible. The ideas proposed in the paper apply to *any* ML model that can be represented with FOL. This encompasses *any* existing ML model, with minor adaptations in case the ML model keeps state.

[10]The abbreviations used relate with the names in the decision tree, and serve for saving space.



(a) Decision tree

IF	$(\mathsf{L} = \mathsf{lng})$	**THEN**	skips	(R1)
IF	$(\mathsf{L} = \mathsf{shrt}) \wedge (\mathsf{T} = \mathsf{flUp}) \wedge (\mathsf{A} = \mathsf{ukwn})$	**THEN**	skips	(R2)
IF	$(\mathsf{L} = \mathsf{shrt}) \wedge (\mathsf{T} = \mathsf{new})$	**THEN**	reads	(R3)
IF	$(\mathsf{L} = \mathsf{shrt}) \wedge (\mathsf{T} = \mathsf{flUp}) \wedge (\mathsf{A} = \mathsf{kwn})$	**THEN**	reads	(R4)

(b) Rule set

$$\mathcal{M}_\pi(\mathsf{L}, \mathsf{T}, \mathsf{A}, \mathsf{W}) \triangleq$$
$$[(\mathsf{L} \vee \neg\mathsf{L} \wedge \mathsf{T} \wedge \neg\mathsf{A}) \rightarrow (\pi = \mathsf{skips})] \wedge$$
$$[(\neg\mathsf{L} \wedge \neg\mathsf{T} \vee \neg\mathsf{L} \wedge \mathsf{T} \wedge \mathsf{A}) \rightarrow (\pi = \mathsf{reads})]$$

(c) Encoding of $\mathcal{M}_\pi$

Figure 1: Running example Poole & Mackworth (2010)

Our goal is to reason about the ML model, i.e. to implement model-based reasoning, so we need to propose a logical representation for the ML model.

Example 2 *For implementing model-based reasoning, we need to develop an encoding in some suitable fragment of FOL* [11]*. 0-place predicates* [12] *are used for* L, T, A *and* W*, as follows. We will associate* $(\mathsf{L} = \mathsf{long})$ *with* L *and* $(\mathsf{L} = \mathsf{short})$ *with* $\neg\mathsf{L}$*. Similarly, we associate* $(\mathsf{T} = \mathsf{new})$ *with* $\neg\mathsf{T}$*, and* $(\mathsf{T} = \mathsf{followUp})$ *with* T*. We associate* $(\mathsf{A} = \mathsf{known})$ *with* A *and* $(\mathsf{A} = \mathsf{unknown})$ *with* $\neg\mathsf{A}$*. Furthermore, we associate* $(\mathsf{W} = \mathsf{home})$ *with* $\neg\mathsf{W}$ *and* $(\mathsf{W} = \mathsf{work})$ *with* W*. An example encoding is shown in Figure 1c. The explicit values of* π *are optional (i.e. propositional values could be used) and serve to illustrate how non-propositional valued could be modeled.*

3 Contrastive vs. Abductive Explanations

Recent work Shih et al. (2018; 2019); Ignatiev et al. (2019a); Darwiche (2020) proposed to relate model-based explanations with prime implicants. All these approaches compute a set of feature values which, if unchanged, are sufficient for the prediction. Thus, one can view such explanations as answering a 'Why?' question: *the prediction is the one given, as long as some selected set of feature values is the one given*. In this paper, such explanations will be referred to as *abductive explanations*, motivated by one of the approaches used for their computation Ignatiev et al. (2019a).

3.1 Defining Abductive Explanations (AXps)

As indicated earlier in the paper, we focus on *local model-based* explanations.

Definition 3 (Abductive Explanation) *Given an instance* τ*, with a prediction* π*, and an ML model represented with a predicate* $\mathcal{M}_\pi$*, i.e.* $\tau \vDash \mathcal{M}_\pi$*, an* abductive explanation *is a minimal subset of literals of* τ*,* $\sigma \subseteq \tau$*, such that* $\sigma \vDash \mathcal{M}_\pi$*.*

Example 3 *With respect to Example 1, let us consider the instance* $(\mathsf{A} = \mathsf{known}, \mathsf{T} = \mathsf{new}, \mathsf{L} = \mathsf{short}, \mathsf{W} = \mathsf{work})$*, which we will represent instead as* $(\mathsf{A}, \neg\mathsf{T}, \neg\mathsf{L}, \mathsf{W})$*, corresponding to prediction* $\pi = \mathsf{reads}$*. By inspection of the decision tree (seeFigure 1a), a possible answer to the 'Why pred.* reads*?' question is:* $\{\neg\mathsf{L}, \neg\mathsf{T}\}$*. In this concrete case we can conclude that this is the only abductive explanation, again by inspection of the decision tree.*

3.2 Defining Contrastive Explanations (CXps)

As Miller (2019b) notes, contrastive explanations are,

> *"sought in response to particular counterfactual cases... That is, people do not ask why event P happened,but rather why event P happened instead of some event Q."*

As a result, we are interested in providing an answer to the question 'Why π and not δ?', where π is the prediction given some instance τ, and δ is some other (desired) prediction.

[11] Depending on the ML problem, more expressive fragments of FOL logic could be considered.

[12] Which in this case are used as propositional variables.

Example 4 *We consider again Example 1, but with the instance specified in Example 3. A possible answer to the question 'Why pred.* reads *and not pred.* skips*??' is* $\{\mathsf{L}\}$*. Indeed, given the input instance* $(\mathsf{A}, \neg\mathsf{T}, \neg\mathsf{L}, \mathsf{W})$*, if the value of feature* L *changes from* short *to* long*, and the value of the other features remains unchanged, then the prediction will change from* reads *to* skips*.*

The following definition of a (local model-based) contrastive explanation captures the intuitive notion of the contrastive explanation discussed in the example above.

Definition 4 (Contrastive Explanation) *Given an instance* τ*, with a prediction* π*, and an ML model represented by a predicate* $\mathcal{M}_\pi$*, i.e.* $\tau \vDash \mathcal{M}_\pi$*, a* contrastive explanation *is a minimal subset of literals of* τ*,* $\rho \subseteq \tau$*, such that* $\tau \setminus \rho \nvDash \mathcal{M}_\pi$*.*

This definition means that, there is an assignment to the features with literals in ρ, such that the prediction differs from π. Observe that a CXp is defined to answer the following (more specific) question 'Why (pred. π and) not $\neg\pi$?'. The more general case of answering the question 'Why (pred. π and) not δ?' will be analyzed later.

3.3 Relating Abductive & Contrastive Explanations

The previous section proposed a rigorous, model-based, definition of contrastive explanation. Given this definition, one can think of developing dedicated algorithms that compute CXps using a decision procedure for the logic used for representing the ML model. Instead, we adopt a simpler approach. We build on a fundamental result from model-based diagnosis Reiter (1987) (and more generally for reasoning about inconsistency Birnbaum & Lozinskii (2003); Bailey & Stuckey (2005)) and demonstrate a similar relationship between AXps and CXps. In turn, this result reveals a variety of novel algorithms for computing CXps, but also offers ways for enumerating both CXps and AXps.

Local Abductive Explanations (AXps). Consider a set of feature values τ, s.t. the predicion is π, for which the notation $\tau \vDash \mathcal{M}_\pi$ is used. We will use the equivalent statement, $\tau \wedge \neg\mathcal{M}_\pi \vDash \bot$. Thus,

$$\tau \wedge \neg\mathcal{M}_\pi \tag{3}$$

is inconsistent, with the background knowledge being $\mathcal{B} \triangleq \neg\mathcal{M}_\pi$ and the relaxable clauses being $\mathcal{R} \triangleq \tau$. As proposed in Shih et al. (2018); Ignatiev et al. (2019a), a (local abductive) explanation is a subset-minimal set σ of the literals in τ, such that, $\sigma \wedge \neg\mathcal{M}_\pi \vDash \bot$. Thus, σ denotes a subset of the example's input features which, no matter the other feature values, ensure that the ML model predicts π. Thus, any MUS of equation 3 is a (local abductive) explanation for $\mathbb{M}$ to predict π given τ.

Proposition 1 *Local model-based abductive explanations are MUSes of the pair* $(\mathcal{B}, \mathcal{R})$*,* $\tau \wedge \neg\mathcal{M}_\pi$*, where* $\mathcal{R} \triangleq \tau$ *and* $\mathcal{B} \triangleq \neg\mathcal{M}_\pi$*.*

Example 5 *Consider the ML model from Example 1, the encoding from Example 2, and the instance* $\{\mathsf{A}, \neg\mathsf{T}, \mathsf{L}, \neg\mathsf{W}\}$*, with prediction* $\pi = \mathsf{skips}$ *(wrt Figure 1, we replace* $\mathsf{skips} = \mathsf{skips}$ *with* $\mathbf{true}$ *and* $\mathsf{skips} = \mathsf{reads}$ *with* $\mathbf{false}$*). We can thus confirm that* $\tau \vDash \mathcal{M}_\pi$*. We observe that the following holds:*

$$\mathsf{A} \wedge \neg\mathsf{T} \wedge \mathsf{L} \wedge \neg\mathsf{W} \vDash [(\mathsf{L} \vee \neg\mathsf{L} \wedge \mathsf{T} \wedge \neg\mathsf{A}) \rightarrow \mathbf{true}] \wedge [(\neg\mathsf{L} \wedge \neg\mathsf{T} \vee \neg\mathsf{L} \wedge \mathsf{T} \wedge \mathsf{A}) \rightarrow \mathbf{false}] \tag{4}$$

which can be rewritten as,

$$\mathsf{A} \wedge \neg\mathsf{T} \wedge \mathsf{L} \wedge \neg\mathsf{W} \wedge [(\mathsf{L} \vee \neg\mathsf{L} \wedge \mathsf{T} \wedge \neg\mathsf{A}) \wedge \neg\mathbf{true}] \vee [(\neg\mathsf{L} \wedge \neg\mathsf{T} \vee \neg\mathsf{L} \wedge \mathsf{T} \wedge \mathsf{A}) \wedge \neg\mathbf{false}] \tag{5}$$

It is easy to conclude that equation 5 is inconsistent. Moreover, $\sigma = (\mathsf{L})$ *denotes an MUS of equation 5 and denotes one abductive explanation for why the prediction is* skips *for the instance* τ*.*

Local Contrastive Explanations (CXps). Suppose we compute instead an MCS ρ of equation 3, with $\rho \subseteq \tau$. As a result, $\bigwedge_{l \in \tau \setminus \rho}(l) \wedge \neg\mathcal{M}_\pi \nvDash \bot$ holds. Hence, assigning feature values to the inputs of the ML model is consistent with a prediction that is *not* π, i.e. a prediction of some value other than π. Observe that ρ is a subset-minimal set of literals which causes $\tau \setminus \rho \wedge \neg\mathcal{M}_\pi$ to be satisfiable, with any satisfying assignment yielding a prediction that is not π.

Proposition 2 *Local model-based contrastive explanations are MCSes of the pair* $(\mathcal{B}, \mathcal{R})$*,* $\tau \wedge \neg\mathcal{M}_\pi$*, where* $\mathcal{R} \triangleq \tau$ *and* $\mathcal{B} \triangleq \neg\mathcal{M}_\pi$*.*

Example 6 *From equation 3 and equation 5 we can also compute* $\rho \subseteq \tau$ *such that* $\tau \setminus \rho \wedge \neg\mathcal{M}_\pi \nvDash \bot$*. For example* $\rho = (\mathsf{L})$ *is an MCS of equation 5* [13]*. Thus, from* $\{\mathsf{A}, \neg\mathsf{T}, \neg\mathsf{W}\}$ *we can get a prediction other than* skips*, by considering feature value* $\neg\mathsf{L}$*.*

[13] Although in general not the case, in Example 5 and Example 6 an MUS of size 1 is also an MCS of size 1.

Duality Among Explanations. Given the results above, and the hitting set duality between MUSes and MCSes Reiter (1987); Birnbaum & Lozinskii (2003), we have the following.

Theorem 1 *AXps are MHSes of CXps and vice-versa.*

Proof. Immediate from Definition 3, Definition 4, Proposition 1, Proposition 2, and Theorem 4.4 and Corollary 4.5 of Reiter (1987). □

Proposition 1, Proposition 2, and Theorem 1 can now serve to exploit the vast body of work on the analysis of inconsistent formulas for computing both contrastive and abductive explanations and, arguably more importantly, to enumerate explanations. Existing algorithms for the extraction and enumeration of MUSes and MCSes require minor modications to be applied in the setting of AXps and CXps (The resulting algorithms are briefly summarized in Appendix B. Interestingly, a consequence of the duality is that computing an abductive explanation is *harder* than computing a contrastive explanation in terms of the number of calls to a decision procedure Appendix B.).

Discussion. As observed above, the contrastive explanations we are computing answer the question: 'Why (π and) not $\neg\pi$?'. A more general contrastive explanation would be 'Why (π and) not δ, with $\pi \neq \delta$?' Miller (2019b). Note that, since the prediction π is given, we are only interested in changing the prediction to either $\neg\pi$ or δ. We refer to answering the first question as a *basic* contrastive explanation, whereas answering the second question will be referred to as a *targeted* contrastive explanation, and written as CXp_δ. The duality result between AXps and CXps in Theorem 1 applies *only* to basic contrastive explanations. Nevertheless, the algorithms for MCS extraction for computing a basic CXp can also be adapted to computing targeted CXps, as follows. We want a pick of feature values such that the prediction is δ. We start by letting all features to take any value, and such that the resulting prediction is δ. We then iteratively attempt to fix feature values to those in the given instance, while the prediction remains δ. This way, the set of literals that change value are a subset-minimal set of feature-value pairs that is sufficient for predicting δ. Finally, there are crucial differences between the duality result established in this section, which targets local explanations, and a recent result Ignatiev et al. (2019b), which targets *global* explanations. Earlier work established a relation between prime implicants and implicates as a way to relate global abductive explanations and so-called counterexamples. In contrast, we delved into the fundamentals of reasoning about inconsistency, concretely the duality between MCSes and MUSes, and established a relation between model-based *local* AXps and CXps.

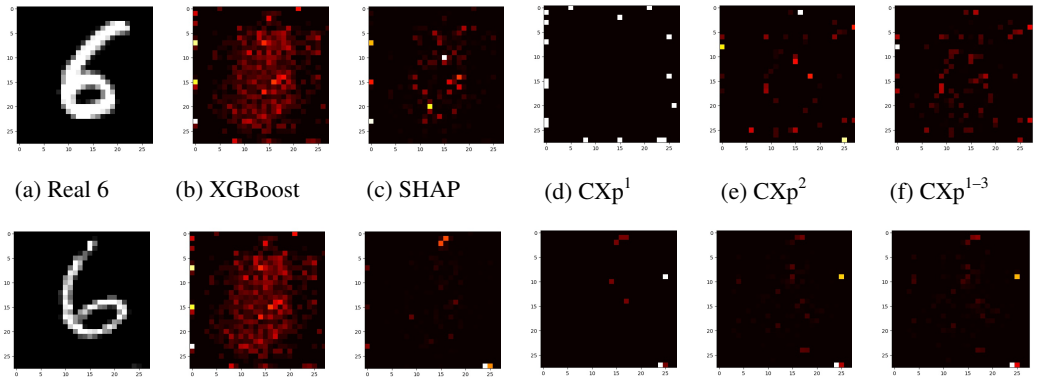
4 Experimental Evaluation

This section details the experimental evaluation to assess the practical feasibility and efficiency of the enumeration of abductive and contrastive explanations for a few real-world datasets, studied in the context of explainability and algorithmic fairness. To perform the evaluation, we adapt powerful algorithms for enumeration MCSes or MCSes and MUSes to find all abductive and contrastive explanations Bailey & Stuckey (2005); Liffiton & Sakallah (2008); Grégoire et al. (2018); Bendík et al. (2018) [14]. Algorithm 1 and Algorithm 2 in Appendix B show our adaptations of MCS (resp. MCS and MUS) enumeration algorithms to the enumeration of CXps (resp. AXps and CXps).

Enumeration of CXps. Our experiments demonstrate a novel, unexpected practical use case of CXps enumeration algorithms. We show that our method gives a *new fine-grained view* on both global and local standard explanations extracted from ML models. We conduct two sets of experiments. The first experiment, called "real vs fake", distinguishes real from fake images. A dataset contains two classes of images: (a) original MNIST digits and (b) fake MNIST digits produced by a standard DCGAN model Radford et al. (2016) (see Figure 2a and Figure 2g for typical examples). The second experiment, called "3 vs 5 digits", uses a dataset that contains digits "3" and "5" from the standard MNIST dataset (discussed in Appendix C.1). Next, we discuss the results of the "real vs fake" experiment in details (Figure 2). For "real vs fake", we train an XGBoost model Chen & Guestrin (2016) with 100 trees of depth 6 (accuracy 0.85/0.80 on train/test sets). We quantized images so that each pixel takes a value between 0 and 15, image pixels are categorical features in the model.

Global and local explainers. We start by discussing our results on a few samples (Figure 2a and Figure 2g). First, we extract important features provided by XGBoost. As these features are *global* for the model, they are the same for all inputs (Figure 2b and Figure 2h are identical for real and fake images). Figure 2b shows that these important features are *no very informative* for this dataset as these

[14] The prototype and the experimental setup are available at `https://github.com/xdual/xdual`.



(a) Real 6 (b) XGBoost (c) SHAP (d) CXp^1 (e) CXp^2 (f) CXp$^{1–3}$

(g) Fake 6 (h) XGBoost (i) SHAP (j) CXp^3 (k) CXp^4 (l) CXp$^{3–5}$

Figure 2: The 'real vs fake' images. The first row shows results for the real image 6; the second – results for the fake image 6. The first column shows examples of inputs; the second – heatmaps of XGBoost's important features; the third – heatmaps of SHAP's explanation. Last three columns show heatmaps of CXp of different cardinality. The brighter pixels are more influential features.

pixels form a blob of pixels that cover an image. Then we compute an image-specific explanation using the standard explainer SHAP (see Figure 2c for the real image and Figure 2i for the fake image). SHAP explanations are more focused on specific parts of images compared to XGBoost. However, it is still not easy to gain insights about the classifier behaviour from them either. For example, both XGBoost and SHAP distinguish some edge and middle pixels as key pixels (the bright pixels are more important) but it is not clear why these are important pixels.

CXps enumeration approach. Given all CXps of size k, we plot a heatmap of occurrences of each pixel in these CXps of size k. Let us focus on the first row with the real 6. For example, CXp^2 (Figure 2e) shows the heatmap of CXps of size two for the real 6. As we mentioned above, both XGBoost and SHAP hint that a few edge pixels are important. Our method can explain *why* this is the case. Consider the heatmap CXp^1 at Figure 2d that shows all CXps of size one for the real 6. It shows that most of important pixels of XGBoost and SHAP are actually CXps of size one. This means that *it is sufficient to change a single pixel value to some other value to obtain a different prediction*. This insight is an important explanation of the classifier behaviour for the user. Now, consider the heatmap CXp^2 at Figure 2e of CXps of size two. It overlaps a lot with SHAP important pixels in the middle of the image explaining why these are important. Only a *pair* of these pixels can be changed to get a different prediction. Moreover, that our results reveal an interesting observation. DCGAN generates images with a few gray edges pixels (see Figure 4 in Appendix. Indeed, some of them have several edge pixels in gray.) This 'defect' does not happen often for real MNIST images. Therefore, the classifier 'hooks' on this issue to classify an image as fake.

A correlation between CXps and SHAP's important features. To qualitatively measure our observations on correlation between key features of CXps and SHAP, we conducted the same experiment as above on 100 random images and measured the correlation between CXps and SHAP features. First, we compute a set T of pixels that is the union of the first (top) 100 smallest size CXps. On average, we have 60 pixels in T. Note that the average 60 pixels represent a small fraction (7%) of the total number of pixels. Then we find a set S of $|T|$ SHAP pixels with highest absolute weights. Finally, we compute $corr = |S \cap T|/|S|$ as the correlation measure. Note that $corr = 0.4$ on average, i.e. our method hits 40% of best SHAP features. As the chances of two tools independently hitting the same pixel (out of 784) are quite low, the fact that 40% of $|S|$ are picked indicates a significant correlation.

Enumeration of CXps and AXps. We consider the six well-known and publicly available datasets. Three of them were previously studied in Ribeiro et al. (2018) in the context of heuristic explanation approaches, namely, Anchor Ribeiro et al. (2018) and LIME Ribeiro et al. (2016), including *Adult*, *Lending*, and *Recidivism*. Appendix C.2 provides a detailed explanation of datasets and our implementation. A prototype implementing is an adaptation of Liffiton & Sakallah (2008) targeting the enumeration of either (1) all abductive or (2) all contrastive explanations was created. In the experiment, the prototype implementation is instructed to enumerate all abductive explanations. The prototype is able to deal with tree ensemble models trained with XGBoost Chen & Guestrin (2016). Given a dataset, we trained an XGBoost model containing 50 trees per class, each tree having depth

	Dataset					
	Adult	**Lending**	**Recidivism**	**Compas**	**German**	**Spambase**
# of instances	5579.0	4414.0	3696.0	778.0	1000.0	2344.0
total time (sec.)	7666.9	443.8	3688.0	78.4	16 943.2	6859.2
minimal time (sec.)	0.1	0.0	0.1	0.0	0.2	0.1
average time (sec.)	1.4	0.1	1.0	0.1	16.9	2.9
maximal time (sec.)	13.1	0.8	8.9	0.5	193.0	23.1
total oracle calls	492 990.0	69 653.0	581 716.0	21 227.0	748 164.0	176 354.0
minimal oracle calls	14.0	11.0	17.0	13.0	23.0	12.0
average oracle calls	88.4	15.8	157.4	27.3	748.2	75.2
maximal oracle calls	581.0	73.0	1426.0	134.0	7829.0	353.0
total # of AXps	52 137.0	8105.0	60 688.0	1931.0	59 222.0	18 876.0
average # of AXps	9.4	1.8	16.4	2.5	59.2	8.1
average AXp size	5.3	1.9	6.4	3.8	7.5	4.6
total # of CXps	66 219.0	8663.0	77 784.0	3558.0	66 781.0	24 774.0
average # of CXps	11.9	2.0	21.1	4.6	66.8	10.6
average CXp size	2.4	1.4	2.6	1.5	3.6	2.3

Table 1: Results of the computational experiment on enumeration of AXps and CXps.

3. (Further increasing the number of trees per class and also increasing the maximum depth of a tree did not result in a significant increase of the models' accuracy on the training and test sets for the considered datasets.) All abductive explanations for every instance of each of the six datasets were exhaustively enumerated using the duality-based approach (Algorithm 2 in Appendix B). This resulted in the computation of all contrastive explanations as well).

Evaluation results. Table 1 shows the results. There are several points to make. First, although it seems computationally expensive to enumerate all explanations for a data instance, it can still be achieved effectively for the medium-sized models trained for all the considered datasets. This may on average require from a few dozen to several hundred of oracle calls per data instance (in some cases, the number of calls gets up to a few thousand). Also observe that enumerating all explanations for an instance takes from a fraction of a second to a couple of seconds on average. These results demonstrate that our approach is practical.

Second, the total number of AXps is typically lower than the total number of their contrastive counterparts. The same holds for the average numbers of abductive and contrastive explanations per data instance. Third and finally, AXps for the studied datasets tend to be larger than contrastive explanations. The latter observations imply that contrastive explanations may be preferred from a user's perspective, as the smaller the explanation is the easier it is to interpret for a human decision maker. (Furthermore, although it is not shown in Table 1, we noticed that in many cases contrastive explanations tend to be of size 1, which makes them ideal to reason about the behaviour of an ML model.) On the other hand, exhaustive enumeration of contrastive explanations can be more time consuming because of their large number.

Summary of results. We show that CXps enumeration gives us an insightful understanding of a classifier's behaviour. First, even in cases when we cannot enumerate all of CXps to compute AXps by duality, we can still draw some conclusions, e.g. CXps of size one are exactly features that occur in all AXps. Next, we clearly demonstrate the feasibility of the duality-based exhaustive enumeration of both AXps and CXps for a given data instance using a more powerful algorithm that performs enumeration of AXps and CXps.

5 Conclusions

This paper studies local model-based abductive and contrastive explanations. Abductive explanations answer 'Why?' questions, whereas contrastive explanations answer 'Why Not?' questions. Moreover, the paper relates explanations with the analysis of inconsistent theories, and shows that abductive explanations correspond to minimal unsatisfiable subsets, whereas contrastive explanations can be related with minimal correction subsets. As a consequence of this result, the paper exploits a well-known minimal hitting set relationship between MUSes and MCSes Reiter (1987); Birnbaum & Lozinskii (2003) to reveal the same relationship between abductive and contrastive explanations. In addition, the paper exploits known results on the analysis of inconsistent theories, to devise algorithms for extracting and enumerating abductive and contrastive explanations.

REFERENCES

P. Achinstein. *The Nature of Explanation*. Oxford University Press, 1980.

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. doi: 10.1109/ACCESS.2018.2870052.

José M. Alonso, Ciro Castiello, and Corrado Mencar. A bibliometric analysis of the explainable artificial intelligence research field. In *IPMU*, pp. 3–15, 2018. doi: 10.1007/978-3-319-91473-2_1.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. `http://tiny.cc/dd7mjz`, 2016.

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *AAMAS*, pp. 1078–1088, 2019.

Fahiem Bacchus and George Katsirelos. Using minimal correction sets to more efficiently compute minimal unsatisfiable sets. In *CAV*, pp. 70–86, 2015.

James Bailey and Peter J. Stuckey. Discovery of minimal unsatisfiable subsets of constraints using hitting set dualization. In *PADL*, pp. 174–186, 2005.

Jaroslav Bendík, Ivana Cerná, and Nikola Benes. Recursive online enumeration of all minimal unsatisfiable subsets. In *ATVA*, pp. 143–159, 2018.

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pp. 1, 2017.

Elazar Birnbaum and Eliezer L. Lozinskii. Consistent subsets of inconsistent systems: structure and behaviour. *J. Exp. Theor. Artif. Intell.*, 15(1):25–46, 2003.

Alessio Bonfietti, Michele Lombardi, and Michela Milano. Embedding decision trees and random forests in constraint programming. In *CPAIOR*, pp. 74–90, 2015.

S. Bromberger. An approach to explanation. In R. Butler (ed.), *Analytical Philsophy*, pp. 72–105. Oxford University Press, 1962.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *KDD*, pp. 785–794. ACM, 2016.

Adnan Darwiche. Three modern roles for logic in AI. *CoRR*, abs/2004.08599, 2020. URL `https://arxiv.org/abs/2004.08599`.

Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. *CoRR*, abs/2002.09284, 2020. URL `https://arxiv.org/abs/2002.09284`.

Leonardo Mendonça de Moura and Nikolaj Bjørner. Z3: an efficient SMT solver. In *TACAS*, pp. 337–340, 2008.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Pai-Shun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NIPS*, pp. 590–601, 2018.

Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable artificial intelligence: A survey. In *MIPRO*, pp. 210–215, 2018. doi: 10.23919/MIPRO.2018.8400040.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

FairML. Auditing black-box predictive models. `http://tiny.cc/6e7mjz`, 2016.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pp. 259–268. ACM, 2015.

Alexander Felfernig, Monika Schubert, and Christoph Zehentner. An efficient diagnosis algorithm for inconsistent constraint sets. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 26:53 – 62, 01 2012. doi: 10.1017/S0890060411000011.

Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On algorithmic fairness, discrimination and disparate impact. 2015.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *FAT*, pp. 329–338. ACM, 2019.

Nicholas Frosst and Geoffrey E. Hinton. Distilling a neural network into a soft decision tree. In *CEx@AI*IA*, 2017.

Marco Gario and Andrea Micheli. PySMT: a solver-agnostic library for fast prototyping of SMT-based algorithms. In *SMT Workshop*, 2015.

Google. AI Explainability Whitepaper. `http://tiny.cc/tjz2hz`, 2019.

Éric Grégoire, Yacine Izza, and Jean-Marie Lagniez. Boosting MCSes enumeration. In *IJCAI*, pp. 1309–1315, 2018.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5): 93:1–93:42, 2019. doi: 10.1145/3236009.

Robert R. Hoffman and Gary Klein. Explaining explanation, part 1: Theoretical foundations. *IEEE Intelligent Systems*, 32(3):68–73, 2017. doi: 10.1109/MIS.2017.54.

Robert R. Hoffman, Shane T. Mueller, and Gary Klein. Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems*, 32(4):78–86, 2017. doi: 10.1109/MIS.2017.3121544.

Robert R. Hoffman, Tim Miller, Shane T. Mueller, Gary Klein, and William J. Clancey. Explaining explanation, part 4: A deep dive on deep nets. *IEEE Intelligent Systems*, 33(3):87–95, 2018a. doi: 10.1109/MIS.2018.033001421.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: challenges and prospects. *CoRR*, abs/1812.04608, 2018b.

IBM ILOG. IBM ILOG: CPLEX optimizer 12.10.0. `http://tiny.cc/jh7mjz`, 2020.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *AAAI*, pp. 1511–1519, 2019a.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. In *NeurIPS*, pp. 15857–15867, 2019b.

Susmit Jha, Tuhin Sahai, Vasumathi Raman, Alessandro Pinto, and Michael Francis. Explaining AI decisions using efficient methods for learning sparse boolean formulae. *J. Autom. Reasoning*, 63 (4):1055–1075, 2019.

Ulrich Junker. QUICKXPLAIN: preferred explanations and relaxations for over-constrained problems. In *AAAI*, pp. 167–172, 2004.

Gary Klein. Explaining explanation, part 3: The causal landscape. *IEEE Intelligent Systems*, 33(2): 83–88, 2018. doi: 10.1109/MIS.2018.022441353.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, pp. 202–207, 1996.

Mark H. Liffiton and Karem A. Sakallah. Algorithms for computing minimal unsatisfiable subsets of constraints. *J. Autom. Reasoning*, 40(1):1–33, 2008.

Mark H. Liffiton, Alessandro Previti, Ammar Malik, and Joao M. Silva. Fast, flexible MUS enumeration. *Constraints*, 21(2):223–250, 2016.

Michele Lombardi, Michela Milano, and Andrea Bartolini. Empirical decision model learning. *Artif. Intell.*, 244:343–367, 2017.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, pp. 4765–4774, 2017.

Tim Miller. Contrastive explanation: A structural-model approach. *CoRR*, abs/1811.03163, 2018. URL http://arxiv.org/abs/1811.03163.

Tim Miller. "but why?" understanding explainable artificial intelligence. *ACM Crossroads*, 25(3): 20–25, 2019a. doi: 10.1145/3313107.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267: 1–38, 2019b. doi: 10.1016/j.artint.2018.07.007.

Brent D. Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in AI. In *FAT*, pp. 279–288, 2019. doi: 10.1145/3287560.3287574.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018. doi: 10.1016/j.dsp.2017.10.011.

Nina Narodytska, Nikolaj Bjørner, Maria-Cristina V. Marinescu, and Mooly Sagiv. Core-guided minimal correction set and core enumeration. In *IJCAI*, pp. 1353–1361, 2018.

Laurent Perron and Vincent Furnon. Or-tools. URL https://developers.google.com/optimization/.

David Poole and Alan K. Mackworth. *Artificial Intelligence - Foundations of Computational Agents*. Cambridge University Press, 2010. ISBN 978-0-521-51900-7.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06434.

Raymond Reiter. A theory of diagnosis from first principles. *Artif. Intell.*, 32(1):57–95, 1987.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD*, pp. 1135–1144, 2016.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, pp. 1527–1535, 2018.

Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. In Samek et al. (2019), pp. 5–22. ISBN 978-3-030-28953-9. doi: 10.1007/978-3-030-28954-6_1.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer, 2019. ISBN 978-3-030-28953-9. doi: 10.1007/978-3-030-28954-6.

Peter Schmidt and Ann D. Witte. Predicting recidivism in north carolina, 1978 and 1980. *Inter-University Consortium for Political and Social Research*, 1988.

Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In *IJCAI*, pp. 5103–5111, 2018.

Andy Shih, Arthur Choi, and Adnan Darwiche. Compiling bayesian network classifiers into decision graphs. In *AAAI*, pp. 7966–7974, 2019.

Son N. Tran and Artur S. d'Avila Garcez. Deep logic networks: Inserting and extracting knowledge from deep belief networks. *IEEE Trans. Neural Networks Learn. Syst.*, 29(2):246–258, 2018. doi: 10.1109/TNNLS.2016.2603784. URL https://doi.org/10.1109/TNNLS.2016.2603784.

Sicco Verwer, Yingqian Zhang, and Qing Chuan Ye. Auction optimization using regression trees and linear models as integer programs. *Artif. Intell.*, 244:368–395, 2017.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *NLPCC*, pp. 563–574, 2019. doi: 10.1007/978-3-030-32236-6_51.

Appendices

A TAXONOMY

The taxonomy of explanations used in the paper is summarized in Table 2.

Table 2: Taxonomy of ML model explanations used in the paper.

<table>
<tr><td colspan="2"></td><td colspan="3">Instance-</td></tr>
<tr><td colspan="2"></td><td colspan="2">dependent</td><td>independent</td></tr>
<tr><td rowspan="3">ML model-</td><td>agnostic</td><td colspan="2">Heuristic local explanation for π.
Examples: SHAP, LIME, Anchor, etc.</td><td>Heuristic global explanation for π.
Examples: SHAP, LIME (e.g. submodular pick)</td></tr>
<tr><td rowspan="2">based</td><td colspan="2">Rigorous local explanation for π.
Examples:</td><td rowspan="2">Rigorous global explanation for π.
Examples: absolute/global AXps</td></tr>
<tr><td>‘Why π?’
PI- (abductive) explanations (AXps)</td><td>‘Why not $\neg\pi$?’
contrastive (CXps) (our work)</td></tr>
</table>

B EXTRACTING & ENUMERATING EXPLANATIONS

The results of Section 3.3 enable exploiting past work on extracting and enumerating MCSes and MUSes to the setting of contrastive and abductive explanations, respectively. Perhaps surprisingly, there is a stark difference between algorithms for extraction and enumeration of contrastive explanations and abductive explanations. Due to the association with MCSes, one contrastive explanation can be computed with a logarithmic number of calls to a decision procedure Liffiton & Sakallah (2008). Moreover, there exist algorithms for the direct enumeration of contrastive explanations Liffiton & Sakallah (2008). In contrast, abductive explanations are associated with MUSes. As a result, any known algorithm for extraction of one abductive explanation requires at best a linear number of calls to a decision procedure Junker (2004), in the worst-case. Moreover, there is no known algorithm for the direct enumeration of abductive explanations, and so enumeration can be achieved only through the enumeration of contrastive explanations Liffiton & Sakallah (2008); Liffiton et al. (2016); Felfernig et al. (2012).

We adapt state-of-the-art algorithms for the enumeration MUSes and MCSes to find all the abductive and contrastive explanations. Note that as in the case of enumeration of MCSes and MUSes, the enumeration of CXps is comparatively easier than the enumeration of AXps. Algorithm 1 shows our adaptation of MCS enumeration algorithm to the enumeration of CXps Liffiton & Sakallah (2008). Other alternatives Grégoire et al. (2018) could be considered instead. Algorithm 1 finds a CXp, blocks it and finds the next one until no more exists. To extract a single CXp, we can use standard algorithm, e.g. Bailey & Stuckey (2005). In principle, enumeration of AXps can be achieved by computing all CXps and then computing all the minimal hitting sets of all CXps, as proposed in the propositional setting Liffiton & Sakallah (2008). However, there are more efficient alternatives that we can adapt here Bailey & Stuckey (2005); Liffiton et al. (2016); Narodytska et al. (2018); Bendík et al. (2018), Algorithm 2 adapts Liffiton et al. (2016) to the case of computing both AXps and CXps. The algorithm simultaneously searches for AXps and CXps and is based on the hitting set duality.

Algorithm 1 Enumeration of CXps

Function CXPENUM ($\mathcal{M}_\pi$,$\mathcal{C}$, π)
  Input: $\mathcal{M}_\pi$: ML model, $\mathcal{C}$: Input cube, π: Prediction
  Variables: $\mathcal{N}$ and $\mathcal{P}$ defined on the variables of $\mathcal{C}$

```
1  I ← ∅ ;                                          // Block CXps
2  while true do
3      μ ← ExtractCXp(M_π, C, π, I)
4      if μ = ∅ then break;
5      ReportCXp(μ)
6      I ← I ∪ NegateLiteralsOf(μ)
```

Algorithm 2 Enumeration of AXps (and CXps)

Function XPENUM ($\mathcal{M}_\pi$,$\mathcal{C}$, π)
  Input: $\mathcal{M}_\pi$: ML model, $\mathcal{C}$: Input cube, π: Prediction
  Variables: $\mathcal{N}$ and $\mathcal{P}$ defined on the variables of $\mathcal{C}$

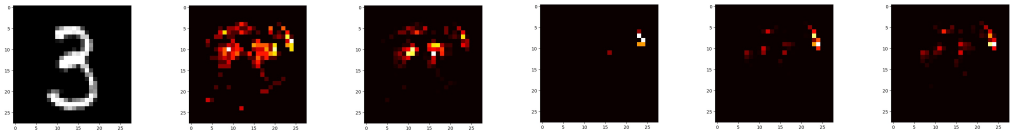
```
1   K = (N, P) ← (∅, ∅) ;                                   // Block AXps & CXps
2   while true do
3       (st_λ, λ) ← FindMHS(P, N) ;                         // MHS of P st N
4       if ¬st_λ then break;
5       (st_ρ, ρ) ← SAT(λ ∧ ¬M_π)
6       if ¬st_ρ then                                       // entailment holds
7           ReportAXp(μ)
8           N ← N ∪ NegateLiteralsOf(μ)
9       else
10          μ ← ExtractCXp(M_π, ρ, π)
11          ReportCXp(ρ)
12          P ← P ∪ UseLiteralsOf(ρ)
```

C ADDITIONAL EXPERIMENTAL RESULTS

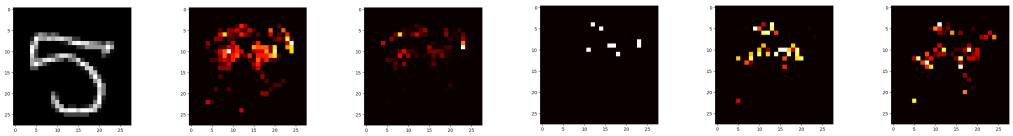
C.1 ENUMERATION OF CXPS

Setup. To perform enumeration of contrastive explanations in our first experiment, we use a constraint programming solver, ORtools Perron & Furnon [15]. To encode the enumeration problem with ORtools we converted scores of XGBoost models into integers keeping 5 digits precision. We enumerate contrastive explanations in the increasing order by their cardinality. This can be done by a simple modification of Algorithm 1 forcing it to return CXps in this order. So, we first obtain all minimum size contrastive explanations, and so on.

Second experiment. Consider our second the "3 vs 5 digits" experiment. We use a dataset that contains digits "3" (class 0) and "5" (class 1) from the standard MNIST (see Figure 3a and Figure 3g for representative samples). XGboost model has 50 trees of depth 3 with accuracy 0.98 (0.97) on train/test sets. We quantized images so that each pixel takes a value between 0 and 15. As before, each pixel corresponds to a feature. So, we have 784 features in our XGBoost model.



(a) Digit 3 (b) XGBoost (c) SHAP (d) CXp^3 (e) CXp^4 (f) CXp$^{3–5}$



(g) Digit 5 (h) XGBoost (i) SHAP (j) CXp^1 (k) CXp^2 (l) CXp$^{1–6}$

Figure 3: Results of the 3 vs 5 digits experiments. The first row shows results for the image 3. The second row shows results for the image 5. The first column shows examples of inputs; the second column shows heatmaps of XGBoost's global important features; the third column shows heatmaps of SHAP's important features. Last three columns show heatmaps of CXp of different cardinality.

Global and local explainers. We start by discussing our results on few random samples (Figure 3a and Figure 3g). First, we obtain the important features from XGBoost. As these features are *global* for the model so they are the same for all inputs (Figure 3b and Figure 3h are identical for 3 and 5

[15]The prototype and the experimental setup are available at `https://github.com/xdual/xdual`.

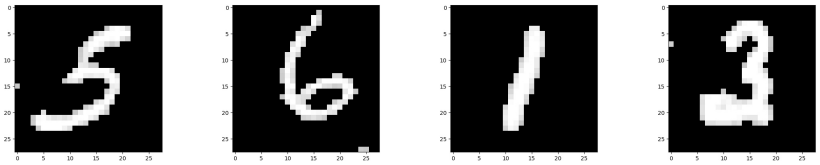


Figure 4: Additional fake images. We reduced values of zero-valued pixels to highlight gray pixels on the edges for some fake images.

images). Figure 2b shows that these important features. The important pixels highlight that the top parts of images are important, which is a plausible high-level explanation of the classifier behavior. Digits 3 and 5 are mostly differ in the top part of the image. However, some pixels are way more important than other and it is hard to understand why.

Next, we compute an image-specific explanation using the standard explainer SHAP (see Figure 3c for the digit 3 and Figure 3c for the digit 5). While SHAP explanations mimic XGBoost important features, they do provide additional insights for the user. Note that both XGBoots and SHAP mark a "belt" of pixels in the upper middle part that as important (bright pixels is the most important pixels).

CXps enumeration approach. We run our enumeration algorithm to produce CXps of increasing cardinality. For each image, we enumerate first 2000 CXps. Given all CXps of size k, we plot a heatmap of occurrences of each pixel in these CXps of size k. Let us focus on the second row with the digit 5. For example, CXp^2 (Figure 3k) shows the heatmap of CXps of size two for the digit 5. As we mentioned above, both XGBoost and SHAP hint that the 'belt' of important pixels in the middle. Again, our method can explain *why* this is the case. Consider the heatmap CXp^1 at Figure 3j. This picture shows all CXps of size one for the digit 5. It reveals that most of important pixels of XGBoost and SHAP are actually CXps of size one. We reiterate that it is sufficient to change a *single* pixel value to some other value to obtain a different prediction. Now, consider the heatmap CXp^{1-6} at Figure 3l. This figure shows 2000 CXps (from size 1 to size 6). It overlaps a lot with SHAP important pixels in the middle of the image. So, these pixels occur in many small size CXps and changing their values leads to misclassification.

Correlation between CXps and SHAP features. To qualitatively measure our observations on correlation between key features of CXps and SHAP, we conducted the same experiment as above on 100 random images and measured the correlation between CXps and SHAP features. First, we compute a set T of pixels that is the union of the first (top) 100 smallest size CXps. On average, we have 38 pixels in T. Note that the average 38 pixels represent a small fraction (5%) of the total number of pixels. Then we find a set S of $|S|$ SHAP pixels with highest absolute weights. Finally, we compute $corr = |S \cap T|/|S|$ as the correlation measure. Note that $corr = 0.6$ on average, i.e. our method hits 60% of best SHAP features. As the chances of two tools independently hitting the same pixel (out of 784) are quite low, the fact that 60% of $|T|$ are picked indicates a significant correlation.

C.2 Enumeration of CXps and AXps

Datasets. The results are obtained on the six well-known and publicly available datasets. Three of them were previously studied in Ribeiro et al. (2018) in the context of heuristic explanation approaches, namely, Anchor Ribeiro et al. (2018) and LIME Ribeiro et al. (2016), including *Adult*, *Lending*, and *Recidivism*. These datasets were processed the same way as in Ribeiro et al. (2018). The *Adult* dataset Kohavi (1996) is originally taken from the Census bureau and targets predicting whether or not a given adult person earns more than $50K a year depending on various attributes, e.g. education, hours of work, etc. The *Lending* dataset aims at predicting whether or not a loan on the Lending Club website will turn out bad. The *Recidivism* dataset was used to predict recidivism for individuals released from North Carolina prisons in 1978 and 1980 Schmidt & Witte (1988). Two more datasets were additionally considered including *Compas* and *German* that were previously studied in the context of the FairML and Algorithmic Fairness projects FairML; Friedler et al. (2015); Feldman et al. (2015); Friedler et al. (2019), an area in which the need for explanations is doubtless. *Compas* is a popular dataset, known Angwin et al. (2016) for exhibiting racial bias of the COMPAS algorithm used for scoring criminal defendant's likelihood of reoffending. The latter dataset is a German credit data (e.g. see Feldman et al. (2015); Friedler et al. (2019)), which given a list of people's attributes classifies them as good or bad credit risks. Finally, we consider the *Spambase*

dataset from the UCI repository Dua & Graff (2017). The main goal is to classify an email as spam or non-spam based on the words that occur in this email. Due to scalability constraints, we preprocessed the dataset to keep ten words per email that were identified as the most influential words by a random forest classifier.

Implementation and Setup. A prototype implementing Algorithm 2 targeting the enumeration of either (1) all abductive or (2) all contrastive explanations was created. In the experiment, the prototype implementation is instructed to enumerate all abductive explanations. (Note that, as was also mentioned before, no matter what kind of explanations Algorithm 2 aims for, all the dual explanations are to be computed as a side effect of the hitting set duality.) The prototype is able to deal with tree ensemble models trained with XGBoost Chen & Guestrin (2016). For that purpose, a simple encoding of tree ensembles into satisfiability modulo theories (SMT) was developed. Concretely, the target formulas are in the theory of linear arithmetic over reals (RIA formulas). (Note that encodings of a decision tree into logic are known Bonfietti et al. (2015); Lombardi et al. (2017); Verwer et al. (2017). The final score summations used in tree ensembles can be encoded into RIA formulas.)

Due to the twofold nature of Algorithm 2, it has to deal with (1) implicit hitting set enumeration and (2) entailment queries with SMT. The former part is implemented using the well-known MILP solver CPLEX IBM ILOG. SMT solvers are accessed through the PySMT framework Gario & Micheli (2015), which provides a unified interface to a variety of state-of-the-art SMT solvers. In the experiments, we use Z3 de Moura & Bjørner (2008) as one of the best performing SMT solvers. The conducted experiment was performed in Debian Linux on an Intel Xeon E5-2630 2.60GHz processor with 64GByte of memory.