Delivering Trustworthy AI through Formal XAI

Joao Marques-Silva[1], **Alexey Ignatiev**[2]
[1] IRIT, CNRS, Toulouse, France
[2] Monash University, Melbourne, Australia
joao.marques-silva@irit.fr, alexey.ignatiev@monash.edu

Abstract

The deployment of systems of artificial intelligence (AI) in high-risk settings warrants the need for trustworthy AI. This crucial requirement is highlighted by recent EU guidelines and regulations, but also by recommendations from OECD and UNESCO, among several other examples. One critical premise of trustworthy AI involves the necessity of finding explanations that offer reliable guarantees of soundness. This paper argues that the best known eXplainable AI (XAI) approaches fail to provide sound explanations, or that alternatively find explanations which can exhibit significant redundancy. The solution to these drawbacks are explanation approaches that offer formal guarantees of rigor. These formal explanations are not only sound but guarantee irredundancy. This paper summarizes the recent developments in the emerging discipline of formal XAI. The paper also outlines existing challenges for formal XAI.

1 Introduction

The vital importance of Trustworthy AI is illustrated by recent guidelines, recommendations and regulations put forward by the European Union (EU), the United States government, the Australian government, the OECD and UNESCO (EU 2016; DARPA 2016; HLEG AI 2019, 2020; EU 2021b,a; National Science and Technology Council (US). Select Committee on Artificial Intelligence 2019; Australian Gov. 2021b,a; OECD 2021; UNESCO 2021). Moreover, the forecast application of machine learning in high-risk and safety-critical applications further underscores the importance of reliable explainable AI (XAI) in delivering trustworthy AI (EU 2021a). Unfortunately, most existing XAI approaches exhibit critical limitations, which represent paramount reasons for excluding their deployment in high-risk and safety-critical settings. To illustrate how significant these critical limitations are, we analyze two concrete scenarios.

The Bessie & Clive affair. *Two friends, Bessie and Clive, are having a drink at the local pub. Bessie is thrilled. Her loan application with Bank 001 was just approved. She will soon be living in her dream house. Clive is devastated. His loan application, also with Bank 001, was just declined. He doubts he will ever own a house. Trying to console her friend, Bessie asks of Clive: "But did the bank explain to you why your application was declined?" Clive retorts: "Well, according to the bank's AI, my application was declined because my age range is 30-45 and my income range is 50K-70K." Bessie looks baffled and exclaims: "That's absurd! The bank's AI told me that my loan application was approved for the exact same reasons!" Enraged with the irrationality of Bank 001's AI, Clive, aided by his lawyer friend Bessie, sues all the banks in the land that use the same AI.*

The fictional story above reveals a critical limitation of existing model-agnostic explanations approaches (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2018). Given two samples $\mathbf{s}_1$ and $\mathbf{s}_2$, with different predictions c_1 and c_2, model agnostic approaches may compute an explanation X that is consistent with both $\mathbf{s}_1$ and $\mathbf{s}_2$. As the example illustrates, such explanations offer no indication whatsoever for why the prediction is c_1 for $\mathbf{s}_1$, and c_2 for $\mathbf{s}_2$. Clearly, limitations like the one above contribute not to building trust in the use of ML-enabled systems, but instead further motivate distrust. Recent work demonstrated that situations similar to the scenario above were observed with a frequency of up to 99.7% for some datasets and ML models (Ignatiev, Narodytska, and Marques-Silva 2019c; Ignatiev 2020). Another issue is that model-agnostic approaches may find unrelated explanations for the same instance.

Mrs. Long's long reasons. *Bank 001's branch manager, May B. Long, is delighted with the new interpretable AI for deciding bank loan applications. Unfortunately, Mrs. Long soon realizes that she must curb her expectations about the new AI, since it often offers explanations that are unacceptably and unnecessarily long, and about which loan applicants complain of being both obscure and inept.*

The second fictional story above hints at a critical limitation of so-called *intrinsically interpretable* ML models (Molnar 2020; Rudin 2019), which represent an alternative to model-agnostic explainers, and which have been advocated for high-risk settings (Rudin 2019). Indeed, it has been shown (Izza, Ignatiev, and Marques-Silva 2020; Huang et al. 2021b) that some interpretable models, namely decision trees, may produce unnecessarily conservative (and so unnecessarily complicated) explanations. Existing ex-

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

perimental evidence obtained on DTs (Izza, Ignatiev, and Marques-Silva 2020) confirms that conservative explanations are often observed in more than 80% of the paths, for DTs obtained with state-of-the-art decision tree learners.

Recent years witnessed a number of efforts towards what this paper refers to as *formal XAI* (Shih, Choi, and Darwiche 2018; Ignatiev, Narodytska, and Marques-Silva 2019a; Shih, Choi, and Darwiche 2019; Narodytska et al. 2019; Ignatiev, Narodytska, and Marques-Silva 2019b,c; Darwiche 2020; Ignatiev 2020; Darwiche and Hirth 2020; Audemard, Koriche, and Marquis 2020; Boumazouza et al. 2020; Ignatiev et al. 2020a; Marques-Silva et al. 2020; Izza, Ignatiev, and Marques-Silva 2020; Barceló et al. 2020; Marques-Silva et al. 2021; Izza and Marques-Silva 2021; Ignatiev and Marques-Silva 2021; Asher, Paul, and Russell 2021; Wäldchen et al. 2021; Huang et al. 2021b; Audemard et al. 2021a; Boumazouza et al. 2021; Blanc, Lange, and Tan 2021; Arenas et al. 2021; Darwiche and Marquis 2021; Ignatiev et al. 2022; Huang et al. 2022; Gorji and Rubin 2022). In contrast with other approaches to XAI, which are currently more visible, formal XAI is based on rigorously defined (and so formal) explanations, ensuring a level of rigor that directly correlates with the logic languages used for representing ML models. The main objective of this paper is to provide an account of the emerging field of formal XAI, highlighting its successes, but also being clear about its current limitations.

The paper is organized as follows. Section 2 introduces the notation and definitions used throughout the paper. This section also briefly overviews the best-known approaches for computing explanations, highlighting their limitations. Section 3 introduces formal explanations, summarizes a number of results related with formal explanations, and a number of computational problems of interest. Section 4 covers the advances in computational formal explanations that have recently been reported. Section 5 investigates the weaknesses of formal explanations and possible approaches for overcoming them. Section 7 concludes the paper.

2 Preliminaries

Throughout this paper, we study explanations for classification problems. A classification problem is represented by a 4-tuple $(\mathcal{F}, \mathbb{D}, \mathcal{K}, \kappa)$. $\mathcal{F} = \{1, \ldots, m\}$ represents a set of m features. $\mathbb{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ denotes a set of domains. Each feature $i \in \mathcal{F}$ is associated with a domain $\mathcal{D}_i$, representing the values that can be assigned to the feature. Given $\mathcal{F}$ and $\mathbb{D}$, feature space is defined as $\mathbb{F} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_m$. (Although the examples in the paper consider Boolean features, formal explanations do not impose restrictions on feature domains.) $\mathcal{K} = \{c_1, \ldots, c_K\}$ is a set of classes. Finally, κ is a non-constant classification function mapping points in feature space to classes, $\kappa : \mathbb{F} \to \mathcal{K}$. An instance is a pair $(\mathbf{v}, c)$, such that $\mathbf{v} \in \mathbb{F}$, $c \in \mathcal{K}$, and $c = \kappa(\mathbf{v})$. The classification function represents the operation of a classifier, i.e. a machine learning (ML) model. We can consider neural networks, tree ensembles, decision trees, etc.

Throughout the paper, the simple neural network (NN) shown in Figure 1a is used as the running example. For this example, $\mathcal{F} = \{1, 2\}$, $\mathbb{D} = \{\mathcal{D}_1, \mathcal{D}_2\}$, with $\mathcal{D}_1 = \mathcal{D}_2 = \{0, 1\}$, and so $\mathbb{F} = \{0, 1\}^2$, $\mathcal{K} = \{0, 1\}$, and $\kappa(x_1, x_2) = (\max(x_1 + x_2 - 0.5, 0) > 0)$. It is easy to conclude that the classifier corresponds to the Boolean function $f(x_1, x_2) = x_1 \vee x_2$.

The paper assumes a basic knowledge of propositional and decidable fragments of first order logic (Biere et al. 2021). The notation used is standard.

Non-formal explanations. These can be broadly categorized as local or global (Guidotti et al. 2019; Samek et al. 2019; Molnar 2020; Samek et al. 2021). Local approaches for computing explanations aim at being valid for points in feature space that are close to the target point, whereas global approaches aim at finding an interpretable model that mimics the original ML model. Most often, local (or *post-hoc*) explanations compute either a simpler model, which is easier to understand and analyze, or a set of features that justify the explanation. LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) illustrate approaches that compute simpler models, and Anchors (Ribeiro, Singh, and Guestrin 2018) illustrates approaches that compute sets of features. As argued in Section 1, and since 2019, several works have identified a growing number of shortcomings of non-formal explanation approaches (Ignatiev, Narodytska, and Marques-Silva 2019c; Narodytska et al. 2019; Ignatiev 2020; Slack et al. 2020; Camburu et al. 2019; Dimanov et al. 2020; Izza et al. 2021). Some of the identified shortcomings demonstrate the inadequacy of non-formal explanations in high-risk settings. For example, the issue with explanations illustrated with the abstract loan example outlined in Section 1 is not hypothetical and has been observed for a number of datasets (Ignatiev, Narodytska, and Marques-Silva 2019c; Ignatiev 2020).

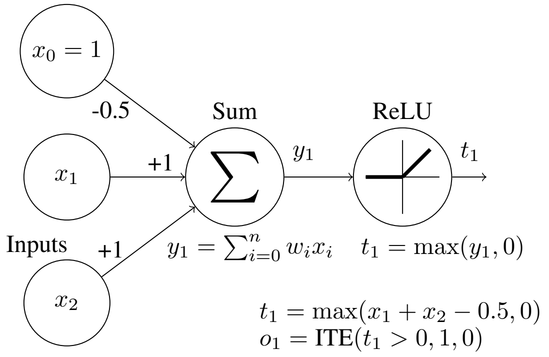
Non-formal explanations exhibit additional important drawbacks. First, for deployed ML models, model agnostic approaches are impractical since these are required to analyze training data (in order to capture the input distribution), which may not be readily available. Similarly, the training data may not rigorously represent the input distribution, or data and/or concept drift may be observed. Thus, in situations where rigor of explanations is paramount, direct access to and the ability to reason about the underlying ML model are crucial requirements.

3 Formal Explanations

We consider an instance $(\mathbf{v}, c)$ and seek explanations for the instance. Concretely, we seek a set of feature-value pairs which are *sufficient* for the prediction, and which are also minimal, meaning a subset- or a cardinality-minimal set. We first propose two categories of formal explanations. Afterwards, we discuss how formal explanations are computed in practice.

3.1 Defining Formal Explanations

The first category of explanation aims to answer a "*Why?*" question. Explanations are represented as conjunctions of literals, relating a feature with a specific or range of values; these represent the chosen *explanation function*, and aim at improving the interpretability of the explanation. Clearly, we


$x_0 = 1$
-0.5
Sum
ReLU
x_1
+1
y_1
t_1
Inputs
+1
x_2
$y_1 = \sum_{i=0}^{n} w_i x_i$
$t_1 = \max(y_1, 0)$
$t_1 = \max(x_1 + x_2 - 0.5, 0)$
$o_1 = \text{ITE}(t_1 > 0, 1, 0)$

(a) NN computing $x_1 \vee x_2$

x_1	x_2	y_1	t_1	o_1
0	0	-0.5	0	0
1	0	0.5	0.5	1
0	1	0.5	0.5	1
1	1	1.5	1.5	1

(b) Truth table

$$\begin{array}{l} x_1 + x_2 - 0.5 = t_1 - s_1 \\ z_1 = 1 \rightarrow t_1 \leq 0 \\ z_1 = 0 \rightarrow s_1 \leq 0 \\ o_1 = (t_1 > 0) \\ x_1, x_2, z_1, o_1 \in \{0, 1\} \\ t_1, s_1 \geq 0 \end{array}$$

(c) Logic representation

$$\begin{array}{l} 1 + 0 - 0.5 = 0.5 - 0 \\ 1 \vee 0.5 \leq 0 \\ 0 \vee 0 \leq 0 \\ 1 = (0.5 > 0) \\ x_1 = 1, x_2 = 0, z_1 = 0, o_1 = 1 \\ t_1 = 0.5, s_1 = 0 \end{array}$$

(d) Instance $(\mathbf{x}, c) = ((1, 0), 1)$

$$\begin{array}{l} 0 + 0 - 0.5 = 0 - 0.5 \\ 0 \vee 0 \leq 0 \\ 1 \vee 0.5 \leq 0 \\ 0 = (0 > 0) \\ x_1 = 0, x_2 = 0, z_1 = 1, o_1 = 0 \\ t_1 = 0, s_1 = 0.5 \end{array}$$

(e) Checking $(x_1, x_2) = (0, 0)$

$$\begin{array}{l} 1 + 1 - 0.5 = 1.5 - 0 \\ 1 \vee 1.5 \leq 0 \\ 0 \vee 0 \leq 0 \\ 1 = (1.5 > 0) \\ x_1 = 1, x_2 = 1, z_1 = 0, o_1 = 1 \\ t_1 = 1.5, s_1 = 0 \end{array}$$

(f) Checking $(x_1, x_2) = (1, 1)$

The running example used in the paper is the (simple) NN shown in Figure 1a. The actual class is picked with an ITE (if-then-else) operator. As can be concluded from the truth table for the classifier, it computes the function $x_1 \vee x_2$ (see Figure 1b). The NN's logic representation (see Figure 1c) is based on earlier work (Fischetti and Jo 2018). For the point $(1, 0)$ in feature space, the prediction is 1. This can easily be checked from the constraints modeling the NN (see Figure 1a). The logic representation is shown to be consistent with the input assignment when the prediction is 1 (see Figure 1d). To compute an AXp, first consider allowing x_1 to take any value. In this case this means allowing x_1 to take value 0 (besides the value 1 it is assigned to). As can be observed, the prediction is allowed to change (actually in this case it is *forced* to change) (see Figure 1e). Hence, the feature 1 must be included in the AXp. In contrast, by changing x_2 from 0 to 1, the prediction cannot change (see Figure 1f). This means that, if the other features remain unchanged, the prediction is 1, no matter the value taken by x_2. Hence, the feature 2 is dropped from the AXp. As a result, the AXp in this case is $\mathcal{X} = \{1\}$.

Figure 1: Complete example with NN

could consider other explanation functions, as long as these were deemed of interest. Informally, an explanation will then be a set of features $\mathcal{X}$ which, if assigned to the values in $\mathbf{v}$, then the prediction is guaranteed to be c, independently of the values assigned to the remaining features in $\mathcal{F} \setminus \mathcal{X}$. Invoking Occam's razor, we want such set of features to be minimal, and consider subset-minimal sets of features. Formally, the condition above can be represented as follows. We want to find a subset-minimal set $\mathcal{X} \subseteq \mathcal{F}$, such that,

$$\forall (\mathbf{x} \in \mathbb{F}). \left[\left(\bigwedge\nolimits_{i \in \mathcal{X}} (x_i = v_i) \right) \rightarrow (\kappa(\mathbf{x}) = c) \right] \qquad (1)$$

This category of explanations is referred to as abductive explanations (Ignatiev, Narodytska, and Marques-Silva 2019a) (AXp's), or PI-explanations (Shih, Choi, and Darwiche 2018), or sufficient reasons (Darwiche and Hirth 2020). In this paper, we will use the acronym AXp. If (1) holds for some set $\mathcal{X} \subseteq \mathcal{F}$, but $\mathcal{X}$ is not necessarily subset-minimal, then we say that $\mathcal{X}$ is a *weak* AXp. Clearly, for an instance $(\mathbf{v}, c)$, $\mathcal{F}$ itself is a weak AXp. Computing cardinality-minimal explanations makes the problem computationally harder (Ignatiev, Narodytska, and Marques-Silva 2019a). Furthermore, existing experimental evidence (Ignatiev, Narodytska, and Marques-Silva 2019a) does not reveal significant reductions in explanations sizes.

As an example, consider a classifier computing the Boolean function $f(x_1, x_2) = x_1 \vee x_2$, and the instance $((1, 0), 1)$. It is intuitive that the abductive *explanation* (for the prediction is $x_1 = 1$, i.e. the prediction will be 1 as long as x_1 is assigned value 1. (Observe that, in this case, the AXp is unique.) Given the proposed definition of AXp above, it is easy to conclude that for $\mathcal{X} = \{1\}$, it is the case that: $\forall (\mathbf{x} \in \mathbb{F}). [(x_1) \rightarrow (x_1 \vee x_2)]$.

Motivated by recent work on understanding the role of explanations (Miller 2019), a different category of explanation aims to answer a "*Why Not?*" question. Here the objective is to understand what needs to be done to change the prediction. Informally, we want to find a subset of features which, if allowed to take some other value, and when the remaining features remain unchanged given their values in $\mathbf{v}$, then the prediction can be changed to a class other than c. As above, we can target subset- or cardinality-minimal explanations, but will preferably discuss subset-minimal explanations. Formally, the condition above can be represented as follows. We want to find a subset-minimal set $\mathcal{Y} \subseteq \mathcal{F}$, such that,

$$\exists(\mathbf{x} \in \mathbb{F}). \left[\left(\bigwedge\nolimits_{i \in \mathcal{F} \setminus \mathcal{Y}} (x_i = v_i)\right) \land (\kappa(\mathbf{x}) \neq c)\right] \quad (2)$$

This category of explanations is referred to as contrastive explanations (Ignatiev et al. 2020b) (CXp's)[1]. If (2) holds for some set $\mathcal{Y} \subseteq \mathcal{F}$, but $\mathcal{Y}$ is not necessarily a CXp, then we say that $\mathcal{Y}$ is a *weak* CXp.

Consider again a classifier representing the Boolean function $f(x_1, x_2) = x_1 \lor x_2$, and the instance $((1, 0), 1)$. It is intuitive that the contrastive *explanation* for the prediction is $x_1 \neq 1$, i.e. for the prediction to change its value, it is the case that the value of x_1 must also be allowed to change. Given the proposed definition of CXp above, it is easy to conclude that for $\mathcal{Y} = \{1\}$, it is the case that: $\exists(\mathbf{x} \in \mathbb{F}). [(\neg x_2) \land \neg(x_1 \lor x_2)]$, which is a true statement, since $x_1 = 0$ (with $x_2 = 0$) cause the prediction to change.

By building on the work of R. Reiter on model-based diagnosis (Reiter 1987), recent work has proved a fundamental duality relationship between AXp's and CXp's (Ignatiev et al. 2020b): *For an instance* $(\mathbf{v}, c)$*, the AXp's are (subset-)minimal hitting sets of the CXp's and vice-versa.* This result is of utmost importance, as it has been instrumental in devising algorithms for enumeration of explanations (Marques-Silva et al. 2021; Huang et al. 2021b; Ignatiev and Marques-Silva 2021). Besides duality between AXp's and CXp's, it can be shown that both (1) and (2) are monotone, respectively on sets $\mathcal{X}$ and $\mathcal{Y}$.

Finally, AXp's can also be formulated *globally* (Ignatiev, Narodytska, and Marques-Silva 2019b), meaning that these are defined independently of a concrete instance. In this case, recent work proved another duality relationship (Ignatiev, Narodytska, and Marques-Silva 2019b), by relating global AXp's with counterexamples (CEx's). This duality result also highlighted connections between explanations and adversarial examples (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015).

3.2 Computing Formal Explanations

Given the formal definitions of explanations proposed in the previous section, we now investigate how automated reasoners can be used for computing explanations in practice. Examples of automated reasoners include Boolean Satisfiability (SAT), Satisfiability Modulo Theories (SMT), or Mixed Integer Linear Programming (MILP). Consistency checking with a reasoner for theory $\mathcal{T}$ on a $\mathcal{T}$-theory formula $\varphi_\mathcal{T}$ is represented by $\textsf{CO}(\varphi_\mathcal{T}; \mathcal{T})$, and denotes whether $\varphi_\mathcal{T}$ has at least one model (given $\mathcal{T}$), i.e. an interpretation that satisfies $\varphi_\mathcal{T}$. For simplicity, the parameterization on $\mathcal{T}$ is omitted, and so we use $\textsf{CO}(\varphi_\mathcal{T})$ instead. These theory reasoners operate on formulas of a suitable logic language. Given some logic formula φ, $[\![\varphi]\!]_\mathcal{T}$ denotes the encoding of φ in a representation suitable for reasoning by a decision oracle for theory $\mathcal{T}$. (For simplicity, we just use $[\![\varphi]\!]$.) As shown below, the computation of formal explanations assumes the existence of a reasoner that decides the satisfiability (or consistency) of a statement expressed in theory $\mathcal{T}$:

$$\textsf{CO}\left([\![\left(\bigwedge\nolimits_{i \in \mathcal{S}} (x_i = v_i)\right) \land (\kappa(\mathbf{x}) \neq c)]\!]\right) \quad (3)$$

(Observe that $\textsf{CO}$ requires some sort of parameterization using $\mathcal{T}$, $\mathcal{F}$, κ, $\mathbf{x}$, and also $c = \kappa(\mathbf{v})$.)

The function $\textsf{CO}$ takes as argument a statement in theory $\mathcal{T}$, and returns one of two values: $\bot$ (or false) if the statement is inconsistent, and $\top$ (or true) if the statement is consistent. Moreover, both (1) and (2) can be decided with calls to $\textsf{CO}$, as shown next. Regarding (1), by double-negating the formula one gets:

$$\neg\exists(\mathbf{x} \in \mathbb{F}). \left[\left(\bigwedge\nolimits_{i \in \mathcal{X}} (x_i = v_i)\right) \land (\kappa(\mathbf{x}) \neq c)\right] \quad (4)$$

By setting $\mathcal{S} = \mathcal{X}$, it becomes clear that (4) holds (and so (1) holds) iff (3) does not hold. Similarly, by setting $\mathcal{S} = \mathcal{F} \setminus \mathcal{Y}$, it is also clear that (2) holds iff (3) also holds.

The computation of a single AXp or a single CXp can be achieved with a greedy algorithm provided a few requirements are met. First, reasoning in theory $\mathcal{T}$ is required to be monotone, i.e. *in*consistency is preserved if constraints are added to a set of constraints, and consistency is preserved if constraints are removed from a set of constraints. Second, for computing one AXp, the predicate to consider is:

$$\begin{aligned} &\mathbb{P}_{\text{axp}}(\mathcal{S}; \mathcal{T}, \mathcal{F}, \kappa, \mathbf{v}) \triangleq \\ &\qquad \neg\,\textsf{CO}\left([\![\left(\bigwedge_{i \in \mathcal{S}} (x_i = v_i)\right) \land (\kappa(\mathbf{x}) \neq c)]\!]\right) \end{aligned} \quad (5)$$

and for computing one CXp, the predicate to consider is:

$$\begin{aligned} &\mathbb{P}_{\text{cxp}}(\mathcal{S}; \mathcal{T}, \mathcal{F}, \kappa, \mathbf{v}) \triangleq \\ &\qquad \textsf{CO}\left([\![\left(\bigwedge_{i \in \mathcal{F} \setminus \mathcal{S}} (x_i = v_i)\right) \land (\kappa(\mathbf{x}) \neq c)]\!]\right) \end{aligned} \quad (6)$$

(Similar to the case of $\textsf{CO}$, $\mathbb{P}_{\text{axp}}$ and $\mathbb{P}_{\text{cxp}}$ are parameterized by $\mathcal{T}$, $\mathcal{F}$, κ, $\mathbf{v}$, and also $c = \kappa(\mathbf{v})$. For simplicity, this parameterization will be left implicit when convenient. Also, the parameterization on $c = \kappa(\mathbf{v})$, given the ones on κ and $\mathbf{v}$.) Moreover, given that (1) and (2) are monotone, then $\mathbb{P}_{\text{axp}}$ and $\mathbb{P}_{\text{cxp}}$ are also monotone with respect to set $\mathcal{S}$.

Algorithm 1 illustrates the computation of one AXp or one CXp. (To prove that the algorithm is sound, the invariants guarantee that the resulting set $\mathcal{S}$ respects the given predicate, i.e. it computes either a weak AXp or a weak CXp. Moreover, monotonicity of both $\mathbb{P}_{\text{axp}}$ and $\mathbb{P}_{\text{cxp}}$ ensures that Algorithm 1 computes a minimal set, respectively an AXp or a CXp.) Algorithm 1 corresponds to the so-called deletion algorithm used for explaining over-constrained sets of constraints. (For reasoning about diagnosis or inconsistent sets

[1] The definition used for CXp is less strict than what is described in earlier work (Miller 2019). However, Miller's definition is easy to accommodate, by fixing the class one is interested in.

Algorithm 1: Finding one AXp/CXp

```
Input: Predicate $\mathbb{P}$, parameterized by $\mathcal{T}$, $\mathcal{F}$, $\kappa$, $\mathbf{v}$
Output: One XP $\mathcal{S}$

procedure oneXP($\mathbb{P}, \mathcal{T}, \mathcal{F}, \kappa, \mathbf{v}$)
    $\mathcal{S} \leftarrow \mathcal{F}$                          ▷ Initialization: $\mathbb{P}(\mathcal{S})$ holds
    for $i \in \mathcal{F}$ do                    ▷ Loop invariant: $\mathbb{P}(\mathcal{S})$ holds
        if $\mathbb{P}(\mathcal{S} \setminus \{i\}; \mathcal{T}, \mathcal{F}, \kappa, \mathbf{v})$ then
            $\mathcal{S} \leftarrow \mathcal{S} \setminus \{i\}$      ▷ Drop $i$ if $\mathbb{P}(\mathcal{S} \setminus \{i\})$ holds
    return $\mathcal{S}$                      ▷ Returned set $\mathcal{S}$: $\mathbb{P}(\mathcal{S})$ holds
```

of constraints, Algorithm 1 can be traced back to the early 90s (Chinneck and Dravnieks 1991). However, the same general algorithm is also used in Valiant's work (Valiant 1984), and some authors (Juba 2016) argue that it is implicit in works from the 19th century (Mill 1843).) Because of monotonicity of $\mathbb{P}$, finding one AXp/CXp can be solved with any algorithm for the problem of finding a minimal set subject to a monotone predicate (MSMP) (Marques-Silva, Janota, and Mencía 2017), and Algorithm 1 is one such example. However, and besides the generalized formulation of the deletion algorithm shown above, other alternatives include the QuickXplain algorithm (Junker 2004) and the Progression algorithm (Marques-Silva, Janota, and Belov 2013). The main difference between these algorithms is the number of times the predicate is checked, and so the number of times consistency of some formula is tested. In the worst-case scenario, all algorithms require a number of predicate tests that grows linearly with the number of features.

4 Progress in Formal Explanations

Formal explanations raise a number of challenges. We discuss two in this section, and postpone presenting a few additional challenges for the next section. One commonly perceived limitation is that one must resort to a logical language suitable for describing the ML model, and this might be unrealistic. As argued in the previous sections, and as shown in the recent work overviewed in this section, this is more of a misconception than a limitation. A second limitation is scalability. Indeed, the initial experimental evidence (Shih, Choi, and Darwiche 2018; Ignatiev, Narodytska, and Marques-Silva 2019a) could hardly be considered encouraging. However, the last couple of years have witnessed a growing number of results on the efficient computation of explanations, both in theory and in practice. This progress is surveyed next.

4.1 Tractable Explanations

Recent years have shown that, for a growing number of families of classifiers, both (1) and (2), but also (5) and (6), and so Algorithm 1, can be solved in polynomial time. The first work targeted the class of Naive Bayes Classifiers (Marques-Silva et al. 2020). This work proposed a polynomial-time algorithm for computing one AXp, and showed that AXp's could be enumerated with polynomial delay (Marques-Silva et al. 2020). A second work studied decision trees (DTs) (Izza, Ignatiev, and Marques-Silva 2020). Although decision trees are in general considered to be interpretable (Rudin 2019; Molnar 2020), recent work has shown that DTs can have paths that are arbitrarily larger (on the number of features) than the number of features in an AXp (Izza, Ignatiev, and Marques-Silva 2020). (Clearly, if some paths in a DT are arbitrarily larger than AXp's consistent with those paths, then it is difficult to argue for the interpretability of DTs, at least when interpretability equates with explanation *succinctness*. The second hypothetical scenario in Section 1 illustrates this issue.) Furthermore, this work showed that AXp's could be computed in polynomial time. In independent work, and also in the case of DTs, it was shown that computing one smallest explanation was NP-hard (Barceló et al. 2020). Later work (Huang et al. 2021c,b) extended the results on DTs, by introducing explanation graphs (XpG's). XpG's allow explaining in polynomial time a wider range of families of classifiers that comprise decision trees, graphs and diagrams, including their multi-valued variants. Furthermore, this work showed that: i) CXp's could also be computed in polynomial time; ii) duality could be used for enumerating AXp's and CXp's; and iii) for DTs the total number of CXp's is polynomial on the number of tree nodes. In the case of DTs, the fact that the number of CXp's is polynomial on the tree size also served to solve in polynomial time the problem of deciding whether a feature is included in some AXp. Clearly, being able to decide *membership* of features in explanations is crucial in assessing fairness, but also in helping human decision makers to understand the impact of features on predictions. Monotonic classifiers is another class for which tractability results were obtained (Marques-Silva et al. 2021), both for computing AXp's and CXp's. These results were further extended in more recent work (Cooper and Marques-Silva 2021). Finally, more recent results showed that classifiers represented with propositional languages (Darwiche and Marquis 2002) can be explained efficiently for a broad class of languages (Huang et al. 2021a, 2022). Concretely, classifiers represented with d-DNNF (or with any strictly more succinct language) can be explained in polynomial time. The same work (Huang et al. 2021a, 2022) also studied general decision functions (GDFs). GDFs associate a boolean function κ_i with each class $c_i \in \mathcal{K}$, and such that the functions $\{\kappa_1, \ldots, \kappa_K\}$ respect two criteria related with computing a total function (i.e. for any point in $\mathbb{F}$ at least one κ_i takes value 1) and ensuring non-overlap among the classifiers (i.e. for no point in feature space there exist two classifiers taking value 1). For GDFs where each boolean function is represented with the propositional language DNNF (or a strictly more succinct language), then one AXp can be computed in polynomial time (Huang et al. 2021a, 2022).

4.2 Efficient Explanations

For a number of additional families of classifiers, recent work showed that the computation of explanations is computationally hard in theory, but that in practice explanations can be efficiently computed, with a performance that even outperforms model-agnostic approaches. A first work proposed encodings for boosted trees (Ignatiev, Narodytska, and Marques-Silva 2019c; Ignatiev 2020) (BTs), with significant

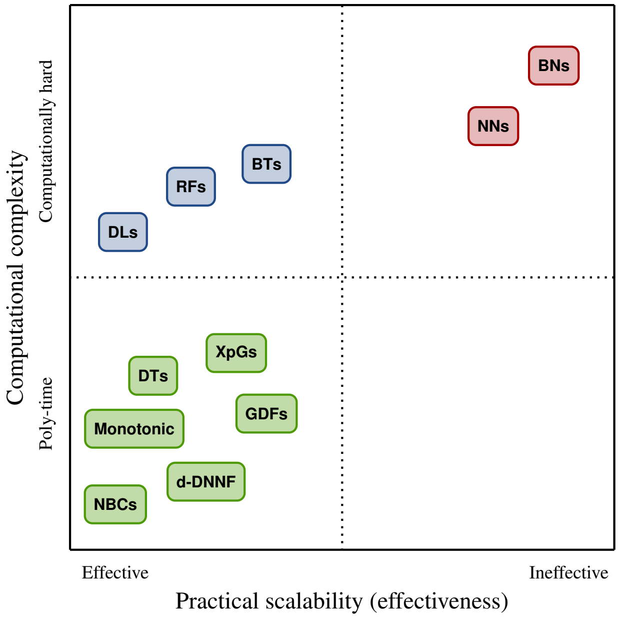


Figure 2: Complexity & practical scalability of finding one (subset-minimal) AXp/CXp

performance improvements reported recently (Ignatiev et al. 2022). More recently, and for random forests (RFs) (Izza and Marques-Silva 2021), the problem of deciding whether a set of features is an AXp was proved to be D^P-complete. However, in practice, the computation of one AXp was shown (Izza and Marques-Silva 2021; Ignatiev et al. 2022) to outperform that of the heuristic method Anchors (Ribeiro, Singh, and Guestrin 2018), allowing the computation of formal explanations for large-size random forests. It should be underscored that this performance difference is observed notably given that Anchors computes a heuristic explanation, without guarantees of soundness or minimality in stark contrast with the guarantees provided by AXp's. A difference encoding for explaining RFs was proposed in more recent work (Boumazouza et al. 2021). Moreover, similar results were obtained for decision lists (Ignatiev and Marques-Silva 2021) (DLs). The progress observed in computing AXp's in recent years is informally summarized in Figure 2. Despite the growing list of ML models for which explanations can be efficiently computed in practice, efficient solutions for neural networks (NNs) and bayesian networks (BNs) remain illusive.

4.3 Approximate Explanations

One limitation of AXp's and CXp's is that there is no control over the size of explanations. Another related limitation is that human decision makers exhibit hard limits on the number of concepts included in explanations (Miller 1956). Recent work proposed δ-relevant inputs (Wäldchen et al. 2021). These can be viewed as a generalization of (smallest) abductive explanations, that allows the prediction not to hold in some points of feature space, as long as the probability of predicting the correct prediction is sufficiently large, and which results in explanations of smaller size. Unfortunately, the complexity of computing δ-relevant inputs is hard for NP^{PP}. As a result, exact computation of δ-relevant inputs appears impractical for most ML models. Nevertheless, recent work showed that by relaxing the definition of δ-relevant set, and for the concrete case of decision trees, then such relaxed (subset-minimal) δ-relevant sets can be computed in polynomial time (Izza et al. 2021). However, a downside of exploiting relaxed δ-relevant sets are the weak guarantees of quality for computed approximate explanations.

4.4 Queries About Explanations

Besides the ability of computing formally defined explanations, another problem of fundamental importance is to be able to decide whether a feature is included in some explanation. For example, if a bank loan application is declined, it would be crucial to decide whether the feature *gender* is contained in some explanation for the decision to decline the loan. (The relationships between (formal) explanations and (formal) *fairness* were reported in recent work (Ignatiev et al. 2020a).) The feature membership (in explanation) problem (FMP) was studied in recent work (Huang et al. 2021b). For functions represented as DNF formulas, (Huang et al. 2021b) showed that the membership problem was Σ_2^P-hard. However, and as mentioned earlier, this same work showed that FMP for decision trees is in P. Given the practical importance of FMP, additional complexity results should be expected in the near future.

Another problem of interest is to be able to enumerate explanations, thus allowing a human decision maker to get a better understanding for the reasons of a prediction. For the case of naive Bayes classifiers, enumeration of abductive explanations can be achieved with polynomial delay (Marques-Silva et al. 2020). More recent work (Marques-Silva et al. 2021; Izza, Ignatiev, and Marques-Silva 2020; Huang et al. 2021b; Ignatiev and Marques-Silva 2021; Huang et al. 2021a, 2022) showed that duality between abductive and contrastive explanations (Ignatiev et al. 2020b) can be used for enumeration of explanations. Enumeration of explanations was also discussed in a more general setting in earlier work (Ignatiev, Narodytska, and Marques-Silva 2019b). Besides queries related with membership and enumeration of explanations, recent work proposed additional queries (Audemard, Koriche, and Marquis 2020; Audemard et al. 2021b).

5 Open Challenges

Despite the rapid progress witnessed with formal XAI, a number of important challenges remain. First, for some relevant classes of classifiers scalability is still an issue. This is the case with neural networks (Ignatiev, Narodytska, and Marques-Silva 2019a), but also with bayesian networks (Shih, Choi, and Darwiche 2019). Recent advances in automated reasoners for NNs (Liu et al. 2021; Katz et al. 2019) are expected to contribute to improving the performance of computing AXp's and CXp's in the case of NNs. Second, the size of formal explanations may be unsuitably large, especially given the cognitive limits of human decision makers. Approximate explanations with δ-relevant sets (Wäldchen et al. 2021), discussed in Section 4.3, offer probabilistic guarantees of rigor, trading off entailment for explanation size. Probabilistic sufficient explanations (Wang, Khosravi, and den Broeck 2021) represent a related effort. There is preliminary work on practical implementations of δ-relevant sets (Izza et al. 2021), for the case of DTs. However, the problem's complexity raises a number of important challenges for the near future. Third, as described in earlier sections, formal explanations assume that any point in feature space is possible. However, in some cases this is not the case. There is preliminary work on constraining the feature space (Gorji and Rubin 2022). Additional work will enable computing explanations by taking input constraints into account. Finally, additional topics include feature aggregation (Ribeiro, Singh, and Guestrin 2016), computing preferred explanations, but also applying formal XAI beyond classification problems.

6 Related Work

Explanations have been comprehensively studied in Artificial Intelligence (Swartout 1977, 1983; Shanahan 1989; Falappa, Kern-Isberner, and Simari 2002; Pérez and Uzcátegui 2003; Amgoud and Prade 2006; Amgoud and Serrurier 2008; Amgoud and Prade 2009; Fan and Toni 2014). Moreover, efforts at formalizing explanations can be traced at least to the mid of the 20th century (Hempel and Oppenheim 1948). Although this paper focuses on AXp's, CXp's, their relationship, and associated computational problems, other formal approaches to explainability have been pursued in recent years (Wolf, Galanti, and Hazan 2019; Amgoud 2021; Liu and Lorini 2021). Other lines of research on explainability (Rago et al. 2020, 2021) are based on formal logic, but are not model-based, and we opt not to categorize them as formal approaches to XAI. The paper opts to cover approaches to formal explainability which have not only seen rapid growth, but are also supported by a stream of practical results. Integration of automated reasoners to improve the quality of results of model-agnostic approaches is also a recent area of research (Shrotri et al. 2022).

7 Conclusions

Non-formal XAI approaches find a growing number of practical uses. Unfortunately, as demonstrated in recent work, their shortcomings make their use untenable in high-risk settings. This paper argues that the only viable alternative for computing explanations in *high-risk situations* involves formal XAI approaches, and the computation of formally-defined explanations. The paper overviews the progress that has been witnessed in formal XAI, it highlights its successes, but it also summarizes its existing limitations, and outlines open areas of research.

Acknowledgments

The work described in this paper results from discussions with several colleagues, including N. Asher, M. Cooper, X. Huang, Y. Izza, N. Narodytska, J. Planes, among others. This work was supported by the AI Interdisciplinary Institute ANITI, funded by the French program "Investing for the Future – PIA3" under Grant agreement no. ANR-19-PI3A-0004, and by the H2020-ICT38 project COALA "Cognitive Assisted agile manufacturing for a Labor force supported by trustworthy Artificial intelligence".

References

Amgoud, L. 2021. Non-monotonic Explanation Functions. In *ECSQARU*, 19–31.

Amgoud, L.; and Prade, H. 2006. Explaining Qualitative Decision under Uncertainty by Argumentation. In *AAAI*, 219–224.

Amgoud, L.; and Prade, H. 2009. Using arguments for making and explaining decisions. *Artif. Intell.*, 173(3-4): 413–436.

Amgoud, L.; and Serrurier, M. 2008. Agents that argue and explain classifications. *Auton. Agents Multi Agent Syst.*, 16(2): 187–209.

Arenas, M.; Baez, D.; Barceló, P.; Pérez, J.; and Subercaseaux, B. 2021. Foundations of Symbolic Languages for Model Interpretability. In *NeurIPS*.

Asher, N.; Paul, S.; and Russell, C. 2021. Fair and Adequate Explanations. In *CD-MAKE*, 79–97.

Audemard, G.; Bellart, S.; Bounia, L.; Koriche, F.; Lagniez, J.; and Marquis, P. 2021a. On the Computational Intelligibility of Boolean Classifiers. In *KR*, 74–86.

Audemard, G.; Bellart, S.; Bounia, L.; Koriche, F.; Lagniez, J.; and Marquis, P. 2021b. On the Computational Intelligibility of Boolean Classifiers. *CoRR*, abs/2104.06172.

Audemard, G.; Koriche, F.; and Marquis, P. 2020. On Tractable XAI Queries based on Compiled Representations. In *KR*, 838–849.

Australian Gov. 2021a. Australia's AU Action Plan. tiny.cc/hy8juz. Accessed: 2021-12-01.

Australian Gov. 2021b. Australia's Artificial Intelligence Ethics Framework. tiny.cc/ey8juz. Accessed: 2021-12-01.

Barceló, P.; Monet, M.; Pérez, J.; and Subercaseaux, B. 2020. Model Interpretability through the lens of Computational Complexity. In *NeurIPS*.

Biere, A.; Heule, M.; van Maaren, H.; and Walsh, T., eds. 2021. *Handbook of Satisfiability*. IOS Press. ISBN 978-1-64368-160-3.

Blanc, G.; Lange, J.; and Tan, L. 2021. Provably efficient, succinct, and precise explanations. In *NeurIPS*.

Boumazouza, R.; Alili, F. C.; Mazure, B.; and Tabia, K. 2020. A Symbolic Approach for Counterfactual Explanations. In *SUM*, 270–277.

Boumazouza, R.; Alili, F. C.; Mazure, B.; and Tabia, K. 2021. ASTERYX: A model-Agnostic SaT-basEd appRoach for sYmbolic and score-based eXplanations. In *CIKM*, 120–129.

Camburu, O.; Giunchiglia, E.; Foerster, J.; Lukasiewicz, T.; and Blunsom, P. 2019. Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods. *CoRR*, abs/1910.02065.

Chinneck, J. W.; and Dravnieks, E. W. 1991. Locating Minimal Infeasible Constraint Sets in Linear Programs. *INFORMS J. Comput.*, 3(2): 157–168.

Cooper, M. C.; and Marques-Silva, J. 2021. On the tractability of explaining decisions of classifiers. In *CP*.

DARPA. 2016. DARPA Explainable Artificial Intelligence (XAI) program. https://www.darpa.mil/program/explainable-artificial-intelligence. Accessed: 2021-12-01.

Darwiche, A. 2020. Three Modern Roles for Logic in AI. In *PODS*, 229–243.

Darwiche, A.; and Hirth, A. 2020. On the Reasons Behind Decisions. In *ECAI*, 712–720.

Darwiche, A.; and Marquis, P. 2002. A Knowledge Compilation Map. *J. Artif. Intell. Res.*, 17: 229–264.

Darwiche, A.; and Marquis, P. 2021. On Quantifying Literals in Boolean Logic and Its Applications to Explainable AI. *J. Artif. Intell. Res.*, 72: 285–328.

Dimanov, B.; Bhatt, U.; Jamnik, M.; and Weller, A. 2020. You Shouldn't Trust Me: Learning Models Which Conceal Unfairness from Multiple Explanation Methods. In *ECAI*, 2473–2480.

EU. 2016. General Data Protection Regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed: 2021-12-01.

EU. 2021a. Artificial Intelligence Act. tiny.cc/wy8juz. Accessed: 2021-12-01.

EU. 2021b. Coordinated Plan on Artificial Intelligence – 2021 Review. https://bit.ly/3hJG2HF. Accessed: 2021-12-01.

Falappa, M. A.; Kern-Isberner, G.; and Simari, G. R. 2002. Explanations, belief revision and defeasible reasoning. *Artif. Intell.*, 141(1/2): 1–28.

Fan, X.; and Toni, F. 2014. On Computing Explanations in Abstract Argumentation. In *ECAI*, 1005–1006.

Fischetti, M.; and Jo, J. 2018. Deep neural networks and mixed integer linear optimization. *Constraints An Int. J.*, 23(3): 296–309.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., *ICLR*.

Gorji, N.; and Rubin, S. 2022. Sufficient Reasons for Classifier Decisions in the Presence of Domain Constraints. In *AAAI*.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5): 93:1–93:42.

Hempel, C. G.; and Oppenheim, P. 1948. Studies in the Logic of Explanation. *Philosophy of science*, 15(2): 135–175.

HLEG AI. 2019. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed: 2021-12-01.

HLEG AI. 2020. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. https://bit.ly/3jAeHds. Accessed: 2021-12-01.

Huang, X.; Izza, Y.; Ignatiev, A.; Cooper, M. C.; Asher, N.; and Marques-Silva, J. 2021a. Efficient Explanations for Knowledge Compilation Languages. *CoRR*, abs/2107.01654.

Huang, X.; Izza, Y.; Ignatiev, A.; Cooper, M. C.; Asher, N.; and Marques-Silva, J. 2022. Tractable Explanations for d-DNNF Classifiers. In *AAAI*.

Huang, X.; Izza, Y.; Ignatiev, A.; and Marques-Silva, J. 2021b. On Efficiently Explaining Graph-Based Classifiers. In *KR*, 356–367.

Huang, X.; Izza, Y.; Ignatiev, A.; and Marques-Silva, J. 2021c. On Efficiently Explaining Graph-Based Classifiers. *CoRR*, abs/2106.01350.

Ignatiev, A. 2020. Towards Trustable Explainable AI. In *IJCAI*, 5154–5158.

Ignatiev, A.; Cooper, M. C.; Siala, M.; Hebrard, E.; and Marques-Silva, J. 2020a. Towards Formal Fairness in Machine Learning. In *CP*, 846–867.

Ignatiev, A.; Izza, Y.; Stuckey, P.; and Marques-Silva, J. 2022. Using MaxSAT for Efficient Explanations of Tree Ensembles. In *AAAI*.

Ignatiev, A.; and Marques-Silva, J. 2021. SAT-Based Rigorous Explanations for Decision Lists. In *SAT*, 251–269.

Ignatiev, A.; Narodytska, N.; Asher, N.; and Marques-Silva, J. 2020b. From Contrastive to Abductive Explanations and Back Again. In *AI*IA*, 335–355.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019a. Abduction-Based Explanations for Machine Learning Models. In *AAAI*, 1511–1519.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019b. On Relating Explanations and Adversarial Examples. In *NeurIPS*, 15857–15867.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019c. On Validating, Repairing and Refining Heuristic ML Explanations. *CoRR*, abs/1907.02509.

Izza, Y.; Ignatiev, A.; and Marques-Silva, J. 2020. On Explaining Decision Trees. *CoRR*, abs/2010.11034.

Izza, Y.; Ignatiev, A.; Narodytska, N.; Cooper, M. C.; and Marques-Silva, J. 2021. Efficient Explanations With Relevant Sets. *CoRR*, abs/2106.00546.

Izza, Y.; and Marques-Silva, J. 2021. On Explaining Random Forests with SAT. In *IJCAI*.

Juba, B. 2016. Learning Abductive Reasoning Using Random Examples. In *AAAI*, 999–1007.

Junker, U. 2004. QUICKXPLAIN: Preferred Explanations and Relaxations for Over-Constrained Problems. In *AAAI*, 167–172.

Katz, G.; Huang, D. A.; Ibeling, D.; Julian, K.; Lazarus, C.; Lim, R.; Shah, P.; Thakoor, S.; Wu, H.; Zeljic, A.; Dill, D. L.; Kochenderfer, M. J.; and Barrett, C. W. 2019. The Marabou Framework for Verification and Analysis of Deep Neural Networks. In *CAV*, 443–452.

Liu, C.; Arnon, T.; Lazarus, C.; Strong, C. A.; Barrett, C. W.; and Kochenderfer, M. J. 2021. Algorithms for Verifying Deep Neural Networks. *Found. Trends Optim.*, 4(3-4): 244–404.

Liu, X.; and Lorini, E. 2021. A Logic for Binary Classifiers and Their Explanation. In *CLAR*, 302–321.

Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *NeurIPS*, 4765–4774.

Marques-Silva, J.; Gerspacher, T.; Cooper, M. C.; Ignatiev, A.; and Narodytska, N. 2020. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In *NeurIPS*.

Marques-Silva, J.; Gerspacher, T.; Cooper, M. C.; Ignatiev, A.; and Narodytska, N. 2021. Explanations for Monotonic Classifiers. In *ICML*, 7469–7479.

Marques-Silva, J.; Janota, M.; and Belov, A. 2013. Minimal Sets over Monotone Predicates in Boolean Formulae. In *CAV*, 592–607.

Marques-Silva, J.; Janota, M.; and Mencía, C. 2017. Minimal sets on propositional formulae. Problems and reductions. *Artif. Intell.*, 252: 22–50.

Mill, J. S. 1843. *A System of Logic, Ratiocinative and Inductive*, volume 1. John W. Parker.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2): 81.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267: 1–38.

Molnar, C. 2020. *Interpretable machine learning*. Lulu.com. https://christophm.github.io/interpretable-ml-book/.

Narodytska, N.; Shrotri, A. A.; Meel, K. S.; Ignatiev, A.; and Marques-Silva, J. 2019. Assessing Heuristic Machine Learning Explanations with Model Counting. In *SAT*, 267–278.

National Science and Technology Council (US). Select Committee on Artificial Intelligence. 2019. The national artificial intelligence research and development strategic plan: 2019 update. https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf.

OECD. 2021. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Accessed: 2021-12-01.

Pérez, R. P.; and Uzcátegui, C. 2003. Preferences and explanations. *Artif. Intell.*, 149(1): 1–30.

Rago, A.; Cocarascu, O.; Bechlivanidis, C.; Lagnado, D. A.; and Toni, F. 2021. Argumentative explanations for interactive recommendations. *Artif. Intell.*, 296: 103506.

Rago, A.; Cocarascu, O.; Bechlivanidis, C.; and Toni, F. 2020. Argumentation as a Framework for Interactive Explanations for Recommendations. In *KR*, 805–815.

Reiter, R. 1987. A Theory of Diagnosis from First Principles. *Artif. Intell.*, 32(1): 57–95.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *KDD*, 1135–1144.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI*, 1527–1535.

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

Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C. J.; and Müller, K. 2021. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proc. IEEE*, 109(3): 247–278.

Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; and Müller, K., eds. 2019. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer. ISBN 978-3-030-28953-9.

Shanahan, M. 1989. Prediction is Deduction but Explanation is Abduction. In *IJCAI*, 1055–1060.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A Symbolic Approach to Explaining Bayesian Network Classifiers. In *IJCAI*, 5103–5111.

Shih, A.; Choi, A.; and Darwiche, A. 2019. Compiling Bayesian Network Classifiers into Decision Graphs. In *AAAI*, 7966–7974.

Shrotri, A. A.; Narodytska, N.; Ignatiev, A.; Meel, K.; Marques-Silva, J.; and Vardi, M. 2022. Constraint-Driven Explanations of Black-Box ML Models. In *AAAI*.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *AIES*, 180–186.

Swartout, W. R. 1977. A Digitalis Therapy Advisor with Explanations. In *IJCAI*, 819–825.

Swartout, W. R. 1983. XPLAIN: A System for Creating and Explaining Expert Consulting Programs. *Artif. Intell.*, 21(3): 285–325.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In Bengio, Y.; and LeCun, Y., eds., *ICLR*.

UNESCO. 2021. Draft Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000374266. Accessed: 2021-12-01.

Valiant, L. G. 1984. A Theory of the Learnable. *Commun. ACM*, 27(11): 1134–1142.

Wäldchen, S.; MacDonald, J.; Hauch, S.; and Kutyniok, G. 2021. The Computational Complexity of Understanding Binary Classifier Decisions. *J. Artif. Intell. Res.*, 70: 351–387.

Wang, E.; Khosravi, P.; and den Broeck, G. V. 2021. Probabilistic Sufficient Explanations. In *IJCAI*, 3082–3088.

Wolf, L.; Galanti, T.; and Hazan, T. 2019. A Formal Approach to Explainability. In *AIES*, 255–261.