Trustable Explainable AI – SAT to the Rescue

Joao Marques-Silva
ICREA & University of Lleida, Spain

Abstract

Explainable Artificial Intelligence (XAI) aims to help human decision makers in understanding the operation of complex AI models. However, many XAI solutions, based on non-symbolic methods, offer no formal guarantees and can produce erroneous results. In contrast, logic-based XAI guarantees the rigor of computed explanations, and this is paramount in high-stakes uses of AI. This talk overviews several flagship applications of Boolean satisfiability (SAT) solvers in reasoning about logic-based explanations.

2012 ACM Subject Classification Theory of computation → Automated reasoning; Theory of computation → Logic and verification; Theory of computation → Constraint and logic programming

Keywords and phrases Explainable AI, Boolean Satisfiability

Digital Object Identifier 10.4230/LIPIcs.SAT.2026.2

Category Invited Talk

Funding *Joao Marques-Silva*: Spanish Government grant PID2023-152814OB-I00.

Acknowledgements Our work on logic-based XAI builds on the input of past collaborators, from the 1990s until the 2020s, at UofM, UofS, UCD, IST/INESC-ID, FCUL, IRIT/CNRS and ICREA/UdL.

1 Logic-Based XAI in a Nutshell

Explainable Artificial Intelligence (XAI) is a cornerstone of trustworthy AI. Most XAI solutions [83, 60, 84, 79] are based on sub-symbolic methods, and thus offer no guarantees of rigor. Examples of critical flaws of these methods have been reported since 2019 [36, 28, 94, 68, 26]. Proposals regarding the use of *interpretable* instead of complex (and so non-interpretable) machine learning (ML) models [85, 86] have also been largely discredited [41, 42, 71, 76]. This section briefly overviews a rigorous alternative to sub-symbolic explainability, namely logic-based (also referred to as formal or symbolic) XAI.

The following definitions are fairly common in the field of logic-based XAI [70, 62, 64]. (Additional detail can be found in these references.) An ML model $\mathcal{M}$ is defined on a set of features $\mathcal{F} = \{1, \ldots, m\}$. Each feature $i \in \mathcal{F}$ has a domain $\mathbb{D}_i$, which can be categorical or ordinal, in which case it can either be discrete or real-valued. The Cartesian product of the domains defines the feature space $\mathbb{F}$. In the case of ML classification, the ML model maps each element of feature space to some class from a set of classes $\mathbb{K} = \{c_1, \ldots, c_K\}$, thus defining a classification function $\kappa : \mathbb{F} \to \mathbb{K}$. κ is required not to be constant, i.e. more than a single class is predicted. Given an *instance* $I = (\mathbf{v}, c)$, with $\mathbf{v} = (v_1, \ldots, v_m)$ and $c = \kappa(\mathbf{v})$, the purpose of XAI is to help a human decision maker to fathom *why* the ML model predicts c. Observe that each v_i is a constant taken from the domain of feature $i \in \mathcal{F}$.

1.1 Abductive & Contrastive Explanations

Answering *Why?* questions. To answer a *Why (the prediction)?* question, logic-based XAI defines abductive explanations [34]. Concretely, a subset $\mathcal{X} \subseteq \mathcal{F}$ is a *weak abductive explanation* (WAXp) if,

$$\forall(\mathbf{x} \in \mathbb{F}).\ (\wedge_{i \in \mathcal{X}}(x_i = v_i)) \rightarrow (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \tag{1}$$

A predicate $\mathsf{WAXp} : 2^{\mathcal{F}} \to \{\top, \bot\}$ is associated with (1). Moreover, an *abductive explanation* (AXp) is any subset-minimal set $\mathcal{X} \subseteq \mathcal{F}$ for which $\mathsf{WAXp}(\mathcal{X})$ holds. The WAXp predicate is monotonically increasing and this is instrumental in practice for computing AXps [34, 62]. Clearly, one can talk about cardinality-minimal AXps, and also about the set of AXps $\mathbb{A}(I)$, for an instance I.

Answering *Why Not?* questions. To answer a *Why Not (some other prediction)?* question, logic-based XAI defines contrastive explanations [33]. Concretely, a subset $\mathcal{Y} \subseteq \mathcal{F}$ is a *weak contrastive explanation* (WCXp) if,

$$\exists(\mathbf{x} \in \mathbb{F}).\ \big(\wedge_{i \in \mathcal{F} \setminus \mathcal{Y}}(x_i = v_i)\big) \land (\kappa(\mathbf{x}) \neq \kappa(\mathbf{v})) \tag{2}$$

A predicate $\mathsf{WCXp} : 2^{\mathcal{F}} \to \{\top, \bot\}$ is associated with (2). Moreover, a *contrastive explanation* (CXp) is any subset-minimal set $\mathcal{Y} \subseteq \mathcal{F}$ for which $\mathsf{WCXp}(\mathcal{Y})$ holds. The WCXp predicate is monotonically increasing and this is instrumental in practice for computing CXps [33, 62]. Clearly, one can talk about cardinality-minimal CXps, and also about the set of CXps $\mathbb{C}(I)$, for an instance I. Moreover, and as pointed out in [59], CXps may differ from counterfactual explanations, despite both being used interchangeably in some works [5, 17, 16].

Building on Reiter's seminal work [82], it has been shown [33] that, for an instance $I = (\mathbf{v}, c)$: (i) $\mathcal{X} \subseteq \mathcal{F}$ is an AXp if and only if it is a minimal hitting set of the set of CXps $\mathbb{C}(I)$; and (ii) $\mathcal{Y} \subseteq \mathcal{F}$ is a CXp if and only if it is a minimal hitting set (MHS) of the set of AXps $\mathbb{A}(I)$. MHS duality between AXps and CXps is instrumental for the enumeration of both AXps and CXps. A feature is *relevant* if it occurs in some element of $\mathbb{A}(I)$ [23, 20, 21]. Given MHS duality between AXps and CXps, a feature is also relevant if it occurs in some element of $\mathbb{C}(I)$. Finally, a feature is AXp-*necessary* (resp. CXp-*necessary*) if it occurs in every element of $\mathbb{A}(I)$ (resp. $\mathbb{C}(I)$).

AXps are MUSes & CXps are MCSes. Since at least Reiter's work [82], it is well-known that one can encode systems (and so also an ML model) $\mathcal{M}$ as a set of first-order logic statements. This observation has been illustrated in several recent works in logic-based XAI [41, 65, 47, 31, 66, 42, 29]. Let T denote the logic encoding of some ML model $\mathcal{M}$. For simplicity, we will assume that T is a propositional formula, represented in CNF, that the features are Boolean and that $\mathbb{K} = \{\top, \bot\}$. The generalization to more expressive logics is straightforward. T is defined on two sets of propositional variables $X \cup Y = \{x_1, \ldots, x_m\} \cup \{y_1, \ldots, y_M\}$, where $X = \{x_1, \ldots, x_m\}$ denotes the propositional variables associated with the features of the ML model $\mathcal{M}$, and $Y = \{y_1, \ldots, y_M\}$ denotes auxiliary variables. One variable, e.g. y_M, is distinguished to denote the class computed by the ML model. For any assignment to the variables in $\{x_1, \ldots, x_m\}$, representing some point $\mathbf{x} \in \mathbb{F}$, there exists at least one assignment to the variables in $\{y_1, \ldots, y_M\}$, such that T is satisfiable, and such that $y_M = \kappa(\mathbf{x})$, and there are no assignments to the variables in $\{y_1, \ldots, y_M\}$, such that T is satisfiable and $y_M \neq \kappa(\mathbf{x})$. Furthermore, we define $h_i \leftrightarrow (x_i \leftrightarrow v_i)$ and $o \leftrightarrow (\kappa(\mathbf{x}) \leftrightarrow \kappa(\mathbf{v}))$. (Converting these constraints to clausal form is straightforward. In this context, $\kappa(\mathbf{v})$ is a constant, and $\kappa(\mathbf{x})$ denotes the value assigned to the corresponding variable y_M in Y.)

Now, we define $R \triangleq [T \wedge_{i=1}^{m} (h_i \leftrightarrow (x_i \leftrightarrow v_i)) \land (o \leftrightarrow (\kappa(\mathbf{x}) \leftrightarrow \kappa(\mathbf{v})))]$, $B \triangleq [R \wedge \neg o]$ and $S \triangleq \wedge_{i=1}^{m} h_i$. We refer to B as the *hard* clauses, and S as the *soft* clauses. (B represents the *background* knowledge one starts from.) It is plain that $B \wedge S$ is unsatisfiable: when the variables in X are assigned the constants defined by $\mathbf{v}$, then the ML model computes $\kappa(\mathbf{v})$, and we are imposing $\neg\kappa(\mathbf{v})$. A minimal unsatisfiable subset (MUS) of the pair (B, S) is a

subset-minimal set $U \subseteq S$ such that $B \land U$ is unsatisfiable. A minimal correction subset (MCS) of the pair (B, S) is a subset-minimal set $C \subseteq S$ such that $B \land S \setminus C$ is satisfiable.[1] Furthermore, $S \setminus C$ is referred to as a maximal satisfiable subset (MSS).

The relationship between the MUSes (resp. MCSes) of the pair (B, S) and AXps (resp. CXps) is well-known [34, 33, 62]. This relationship has important practical consequences, since well-known algorithms for the extraction of MUSes [75, 9], MCSes [67, 78], and their enumeration [58] can be readily exploited in the context of logic-based XAI. This has been underscored in several past works [34, 33, 70, 62, 63, 64].

AXps & logic-based abduction. The connection between AXps and logic-based abduction is also straightforward to derive, despite a recent disconcerting qualm about such a connection [55]. To demonstrate that AXps can be formulated as logic-based abduction, we adopt the notation from earlier reference work on the topic [18]. Using the definitions above, let $H = \{h_1, \ldots, h_m\}$ denote the set of hypotheses, and let $M = o$ denote the manifestation. Furthermore, let $V = \mathsf{vars}(R)$ denote the set of propositional variables used in the definition of R. Finally, the tuple $P = \langle V, H, M, R\rangle$ defines a (propositional) logic-based abduction problem [18], such that an explanation (or solution) for the abduction problem is a set $X \subseteq H$ such that: (i) $R \cup X$ is consistent; and (ii) $R \cup X \models M$. Clearly, for any set $X \subseteq H$, it is the case that $R \cup X$ is consistent, since it suffices to assign to the other x_i variables the values dictated by $\mathbf{v}$. Furthermore, it is also plain that $R \cup X \models M$ corresponds to the set X being sufficient for the prediction. Therefore, abductive explanations are indeed an instantiation of logic-based abduction, when explanations (for the logic-based abduction problem) are restricted to being subset-minimal.

AXps & CXps vs. prime implicants & implicates. AXps were first proposed in early 2019 [34], in the context of arbitrary ML classification, and with the aim of exploiting NP oracles, including SAT and SMT solvers [10]. AXps relate explanations with logic-based abduction. PI-explanations were independently proposed in mid 2018 [90], targeting binary classifiers, and aiming compilation to a canonical representation. PI-explanations represent prime implicants of a restriction of the classifier. The definitions of AXps and PI-explanations are equivalent. Moreover, AXps and PI-explanations capture two key concepts: (i) prediction sufficiency (or entailment, see (1)); and (ii) subset-minimality. WAXps also capture prediction sufficiency. In recent years, the term abductive explanation has found wider application.

From its inception [33], the definition of WCXp (and so also of CXp) uses an existential quantifier (see (2)), i.e. a (W)CXp is obtained by finding some point in feature space such that the prediction changes. Thus, both the definition of AXps and CXps can be computed from the logic representation of the ML model. Hence, while AXps are specific prime implicants of the ML model (given the prediction), CXps are *not* prime implicates. Nevertheless, a different line of work relates CXps with prime implicates of a compiled logic representation [17, 16]. In practice, the compilation of complex ML models raises very significant scalability challenges.

Local, localized & global explanations. Work on sub-symbolic XAI is often described as finding *local* explanations, without a precise definition of what a local explanation signifies. AXps and CXps are referred to as *localized*, in that the explanations are computed with respect to a (local) instance, but hold globally. Furthermore, logic-based explanations have also targeted rigorous definitions of global explanations [35, 8].

[1] The use of set notation in this context is a widely used mild abuse of notation.

1.2 Extensions & Generalizations

The literals of the form $(x_i = v_i)$ used in the definition of AXps/CXps can be generalized to account for other relational operators [23, 42, 52, 46]. To the best of our knowledge, the use of relational operators other than = was first described in [23], with practically efficient algorithms proposed in [42, 46].

AXps represent the deterministic case of probabilistic abductive explanations [91], which have been studied in several works [43, 48, 44, 37, 38]. One challenge of probabilistic abductive explanations is that the definition of the weak case is not monotonic.

For several families of ML models, one AXp/CXp can be computed in polynomial time [41, 65, 66, 23, 14, 22, 15, 12, 7].[2] Polynomial-time enumeration of CXps has been proved for decision trees [23], whereas in the case of Naive Bayes Classifiers, AXps can be enumerated with polynomial delay [65]. For several other families of ML models, it is computationally hard to find one AXp/CXp, e.g. [47, 31, 29, 19]. Nevertheless, sophisticated logic encodings, as well as the use of high-performance MaxSAT solvers, enabled explaining fairly complex ensemble-based ML models. MaxSAT solvers have also been successfully used for computing most general explanations [45] and smallest CXps [40]. The use of incremental MaxSAT solving [80] was shown to be instrumental in this context.

Although the foundations of logic-based XAI assume classification models [90, 34], the core concepts can easily be adapted to the case of regression models. Different works have proposed solutions for handling regression models [4, 6, 64].

Despite the significant progress in logic-based XAI, one key challenge has been the ability to reason about complex neural networks (NNs). The first experiments on computing plain AXps (i.e. those that respect (1) and are subset-minimal) for NNs were presented in 2019 [34], with no observable improvements since then. An alternative line of research has been pursued in recent years. Instead of computing plain AXps, the insight is to compute AXps that hold within some distance of the target instance [92, 39], i.e. the so-called distance-restricted AXps. In turn, this reveals a tight relationship between (distance-restricted) AXps and adversarial robustness [53, 54].

A different approach to explaining complex ML models has been to target model-agnostic instead of model-aware[3] explanations. As shown in recent work [13], model-agnostic explainability can be achieved by the rigorous definition of AXps starting from samples of the ML model behavior, e.g. training data. Sample-based (and so model-agnostic) explainability is not without drawbacks as discussed in [2, 1, 3, 77]. Applications of SAT solvers in sample-based explainability are illustrated in [74].

Finally, the formal guarantees of logic-based XAI only matter if the implemented explainers can be validated for bugs. There has been some work on the certification of logic-based explainers [27] and, more importantly, in demonstrating that publicly available logic explainers can exhibit bugs [24].

1.3 From Feature Selection to Feature Attribution – the Way of Logic

Finding AXps/CXps is tantamount to rigorous XAI by feature selection, i.e. finding sets of features that explain some instance, with a sub-symbolic example being Anchors [83]. However, it is the case that the most widely used sub-symbolic methods of XAI target

[2] Nevertheless, in some cases, e.g. DTs, highly inefficient solutions continue to be proposed [76].

[3] Whereas in model-aware explainability the ML model is encoded into a logical representation, this is not the case in model-agnostic explainability, where reasoning does not require knowledge of the ML model.

feature attribution, i.e. to rank features in terms of their relative importance for some instance. SHAP [60] and LIME [83] are arguably the best-known examples of XAI by feature attribution, where SHAP exemplifies the use of Shapley values [88] in XAI. Recent work [68, 26] demonstrated fundamental flaws in the uses of Shapley values in XAI, proving that the cooperative game used for XAI can mislead human decision makers. This negative result motivated the search of alternatives to Shapley values [11, 93].

Nevertheless, the identified flaws of Shapley values in XAI relate with the cooperative game from which Shapley values are computed, and not with Shapley values themselves. As a result, more recent work [56] proposed a logic-based game for XAI, such that WAXps (or WCXps) are used in the definition of a novel characteristic function. This logic-based game is inspired by the games used in measuring a priori voting power since the 1950s [89]. In practice, researchers have proposed [69, 61] a practically efficient alternative to the tool SHAP, that exploits this novel logic-based game for XAI, but also adopts the computation of WAXps/WCXps from samples [13, 74]. The computation of WAXps in XAI by feature attribution opens new opportunities for exploiting automated reasoners in the rigorous analysis of ML models.

2 SAT & Logic-Based XAI

As underlined in the previous section, logic-based XAI has, since its inception, built extensively on Boolean satisfiability (SAT) solvers and practically efficient uses of SAT solvers as NP oracles.[4] In addition, logic-based XAI has also built on reasoners that extend SAT solvers, e.g. SMT solvers [10]. This talk briefly covers concrete examples of using SAT solvers as NP oracles in the context of logic-based XAI. A few examples are summarized below.

Deciding feature relevancy. Abstraction refinement has become a fundamental technique for solving computational problems that lie beyond NP, with QBF solving representing a well-known example [51, 49, 50]. One case study of exploiting SAT oracles in logic-based XAI has been the problem of deciding feature relevancy [25]. This talk revisits the use of SAT-based abstraction refinement for deciding feature relevancy in the concrete case of random forests. This case study illustrates the importance of problem-specific solutions, in that the use of abstraction refinement has been shown to be orders of magnitude more efficient than a direct encoding to QBF.

Smallest sample-based explanations. In the context of sample-based explainability, one example of a computationally hard function problem is that of computing the smallest abductive explanation [87, 74]. This talk details the application of SAT solvers in computing smallest sample-based abductive explanations.

Enumeration of explanations – the case of decision trees. Another widely used application of SAT oracles has been on the enumeration of MUSes/MCSes, namely the eMUS/MARCO algorithm [57, 81, 58]. In logic-based XAI, eMUS/MARCO-like algorithms have been extensively used for navigating the space of AXps/CXps. This talk details how AXps/CXps are enumerated in the concrete case of decision trees.

[4] There exists extensive bibliography on function problem solving using SAT (or SMT) oracles [10], including MaxSAT and QBF solving. Additional examples include finding subset-minimal sets in different contexts [72, 73], including in cases that lie beyond NP [30, 32].

References

1 Leila Amgoud. Explaining black-box classifiers: Properties and functions. *Int. J. Approx. Reason.*, 155:40–65, 2023. doi:10.1016/J.IJAR.2023.01.004.

2 Leila Amgoud and Jonathan Ben-Naim. Axiomatic foundations of explainability. In *IJCAI*, pages 636–642, 2022. doi:10.24963/IJCAI.2022/90.

3 Leila Amgoud, Martin C. Cooper, and Salim Debbaoui. Axiomatic characterisations of sample-based explainers. In *ECAI*, pages 770–777, 2024. doi:10.3233/FAIA240561.

4 Gilles Audemard, Steve Bellart, Jean-Marie Lagniez, and Pierre Marquis. Computing abductive explanations for boosted regression trees. In *IJCAI*, pages 3432–3441, 2023. doi:10.24963/IJCAI.2023/382.

5 Gilles Audemard, Frédéric Koriche, and Pierre Marquis. On tractable XAI queries based on compiled representations. In *KR*, pages 838–849, 2020. doi:10.24963/KR.2020/86.

6 Gilles Audemard, Jean-Marie Lagniez, and Pierre Marquis. On the computation of contrastive explanations for boosted regression trees. In *ECAI*, pages 1083–1091, 2024. doi:10.3233/FAIA240600.

7 Pablo Barceló, Alexander Kozachinskiy, Miguel Romero, Bernardo Subercaseaux, and José Verschae. Explaining k-nearest neighbors: Abductive and counterfactual explanations. *Proc. ACM Manag. Data*, 3(2):97:1–97:26, 2025. doi:10.1145/3725234.

8 Shahaf Bassan, Guy Amir, and Guy Katz. Local vs. global interpretability: A computational complexity perspective. In *ICML*, pages 3133–3167, 2024. URL: https://proceedings.mlr.press/v235/bassan24a.html.

9 Anton Belov, Inês Lynce, and Joao Marques-Silva. Towards efficient MUS extraction. *AI Commun.*, 25(2):97–116, 2012. doi:10.3233/AIC-2012-0523.

10 Armin Biere, Marijn Heule, Hans van Maaren, and Toby Walsh, editors. *Handbook of Satisfiability - Second Edition*, volume 336 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2021. doi:10.3233/FAIA336.

11 Gagan Biradar, Yacine Izza, Elita Lobo, Vignesh Viswanathan, and Yair Zick. Axiomatic aggregations of abductive explanations. In *AAAI*, pages 11096–11104, 2024. doi:10.1609/AAAI.V38I10.28986.

12 Clément Carbonnel, Martin C. Cooper, and Joao Marques-Silva. Tractable explaining of multivariate decision trees. In *KR*, pages 127–135, 2023. doi:10.24963/KR.2023/13.

13 Martin C. Cooper and Leila Amgoud. Abductive explanations of classifiers under constraints: Complexity and properties. In *ECAI*, pages 469–476, 2023. doi:10.3233/FAIA230305.

14 Martin C. Cooper and Joao Marques-Silva. On the tractability of explaining decisions of classifiers. In *CP*, pages 21:1–21:18, 2021. doi:10.4230/LIPIcs.CP.2021.21.

15 Martin C. Cooper and Joao Marques-Silva. Tractability of explaining classifier decisions. *Artif. Intell.*, 316:103841, 2023. doi:10.1016/J.ARTINT.2022.103841.

16 Adnan Darwiche. Logic for explainable AI. In *LICS*, pages 1–11, 2023. doi:10.1109/LICS56636.2023.10175757.

17 Adnan Darwiche and Chunxi Ji. On the computation of necessary and sufficient explanations. In *AAAI*, pages 5582–5591, 2022. doi:10.1609/AAAI.V36I5.20498.

18 Thomas Eiter and Georg Gottlob. The complexity of logic-based abduction. *J. ACM*, 42(1):3–42, 1995. doi:10.1145/200836.200838.

19 Hao Hu, Alexey Ignatiev, and Joao Marques-Silva. Efficient explanations for rule ensembles. In *CP*, 2026.

20 Xuanxiang Huang, Martin C. Cooper, António Morgado, Jordi Planes, and Joao Marques-Silva. Feature necessity & relevancy in ML classifier explanations. In *TACAS*, pages 167–186, 2023. doi:10.1007/978-3-031-30823-9_9.

21 Xuanxiang Huang, Martin C. Cooper, António Morgado, Jordi Planes, and Joao Marques-Silva. Feature necessity and relevancy in machine learning explanations. *J. Autom. Reason.*, 70(1):4, 2026. doi:10.1007/S10817-026-09748-X.

22 Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, Martin C. Cooper, Nicholas Asher, and Joao Marques-Silva. Tractable explanations for d-DNNF classifiers. In *AAAI*, pages 5719–5728, 2022. doi:10.1609/AAAI.V36I5.20514.

23 Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On efficiently explaining graph-based classifiers. In *KR*, pages 356–367, 2021. doi:10.24963/KR.2021/34.

24 Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. Uncovering bugs in formal explainers: A case study with PyXAI. *CoRR*, abs/2511.03169, 2025. doi:10.48550/arXiv.2511.03169.

25 Xuanxiang Huang, Yacine Izza, and Joao Marques-Silva. Solving explainability queries with quantification: The case of feature relevancy. In *AAAI*, pages 3996–4006, 2023. doi:10.1609/AAAI.V37I4.25514.

26 Xuanxiang Huang and Joao Marques-Silva. On the failings of shapley values for explainability. *Int. J. Approx. Reason.*, 171:109112, 2024. doi:10.1016/J.IJAR.2023.109112.

27 Aurélie Hurault and Joao Marques-Silva. Certified logic-based explainable AI - the case of monotonic classifiers. In *TAP*, pages 51–67, 2023. doi:10.1007/978-3-031-38828-6_4.

28 Alexey Ignatiev. Towards trustable explainable AI. In *IJCAI*, pages 5154–5158, 2020. doi:10.24963/IJCAI.2020/726.

29 Alexey Ignatiev, Yacine Izza, Peter J. Stuckey, and Joao Marques-Silva. Using maxsat for efficient explanations of tree ensembles. In *AAAI*, pages 3776–3785, 2022. doi:10.1609/AAAI.V36I4.20292.

30 Alexey Ignatiev, Mikolás Janota, and Joao Marques-Silva. Quantified maximum satisfiability. *Constraints*, 21(2):277–302, 2016. doi:10.1007/S10601-015-9195-9.

31 Alexey Ignatiev and Joao Marques-Silva. SAT-based rigorous explanations for decision lists. In *SAT*, pages 251–269, 2021. doi:10.1007/978-3-030-80223-3_18.

32 Alexey Ignatiev, António Morgado, and Joao Marques-Silva. Propositional abduction with implicit hitting sets. In *ECAI*, pages 1327–1335, 2016. doi:10.3233/978-1-61499-672-9-1327.

33 Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva. From contrastive to abductive explanations and back again. In *AIxIA*, pages 335–355, 2020. doi:10.1007/978-3-030-77091-4_21.

34 Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *AAAI*, pages 1511–1519, 2019. doi:10.1609/AAAI.V33I01.33011511.

35 Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. In *NeurIPS*, pages 15857–15867, 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/7392ea4ca76ad2fb4c9c3b6a5c6e31e3-Abstract.html.

36 Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On validating, repairing and refining heuristic ML explanations. *CoRR*, abs/1907.02509, 2019. arXiv:1907.02509.

37 Yacine Izza, Xuanxiang Huang, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. On computing probabilistic abductive explanations. *CoRR*, abs/2212.05990, 2022. doi:10.48550/arXiv.2212.05990.

38 Yacine Izza, Xuanxiang Huang, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. On computing probabilistic abductive explanations. *Int. J. Approx. Reason.*, 159:108939, 2023. doi:10.1016/J.IJAR.2023.108939.

39 Yacine Izza, Xuanxiang Huang, António Morgado, Jordi Planes, Alexey Ignatiev, and Joao Marques-Silva. Distance-restricted explanations: Theoretical underpinnings & efficient implementation. In *KR*, 2024. doi:10.24963/KR.2024/45.

40 Yacine Izza, Alexey Ignatiev, Xuanxiang Huang, Peter J. Stuckey, and Joao Marques-Silva. Rigorous explanations for tree ensembles. *CoRR*, abs/2603.29361, 2026. doi:10.48550/arXiv.2603.29361.

41 Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On explaining decision trees. *CoRR*, abs/2010.11034, 2020. arXiv:2010.11034.

42 Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On tackling explanation redundancy in decision trees. *J. Artif. Intell. Res.*, 75:261–321, 2022. doi:10.1613/JAIR.1.13575.

43 Yacine Izza, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. Efficient explanations with relevant sets. *CoRR*, abs/2106.00546, 2021. arXiv:2106.00546.

44 Yacine Izza, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. Provably precise, succinct and efficient explanations for decision trees. *CoRR*, abs/2205.09569, 2022. doi:10.48550/arXiv.2205.09569.

45 Yacine Izza, Alexey Ignatiev, Sasha Rubin, Joao Marques-Silva, and Peter J. Stuckey. Most general explanations of tree ensembles. In *IJCAI*, pages 5463–5471, 2025. doi:10.24963/IJCAI.2025/608.

46 Yacine Izza, Alexey Ignatiev, Peter J. Stuckey, and Joao Marques-Silva. Delivering inflated explanations. In *AAAI*, pages 12744–12753, 2024. doi:10.1609/AAAI.V38I11.29170.

47 Yacine Izza and Joao Marques-Silva. On explaining random forests with SAT. In *IJCAI*, pages 2584–2591, 2021. doi:10.24963/IJCAI.2021/356.

48 Yacine Izza and Joao Marques-Silva. On computing relevant features for explaining NBCs. *CoRR*, abs/2207.04748, 2022. doi:10.48550/arXiv.2207.04748.

49 Mikolás Janota, William Klieber, Joao Marques-Silva, and Edmund M. Clarke. Solving QBF with counterexample guided refinement. In *SAT*, pages 114–128, 2012. doi:10.1007/978-3-642-31612-8_10.

50 Mikolás Janota, William Klieber, Joao Marques-Silva, and Edmund M. Clarke. Solving QBF with counterexample guided refinement. *Artif. Intell.*, 234:1–25, 2016. doi:10.1016/J.ARTINT.2016.01.004.

51 Mikolás Janota and Joao Marques-Silva. Abstraction-based algorithm for 2QBF. In *SAT*, pages 230–244, 2011. doi:10.1007/978-3-642-21581-0_19.

52 Chunxi Ji and Adnan Darwiche. A new class of explanations for classifiers with non-binary features. In *JELIA*, pages 106–122, 2023. doi:10.1007/978-3-031-43619-2_8.

53 Guy Katz, Clark W. Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *CAV*, pages 97–117, 2017. doi:10.1007/978-3-319-63387-9_5.

54 Guy Katz, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljic, David L. Dill, Mykel J. Kochenderfer, and Clark W. Barrett. The marabou framework for verification and analysis of deep neural networks. In *CAV*, pages 443–452, 2019. doi:10.1007/978-3-030-25540-4_26.

55 Faezeh Labbaf, Tomás Kolárik, Martin Blicha, Grigory Fedyukovich, Michael Wand, and Natasha Sharygina. Space explanations of neural network classification. In *CAV*, pages 287–303, 2025. doi:10.1007/978-3-031-98682-6_15.

56 Olivier Létoffé, Xuanxiang Huang, and Joao Marques-Silva. Towards trustable SHAP scores. In *AAAI*, pages 18198–18208, 2025. doi:10.1609/AAAI.V39I17.34002.

57 Mark H. Liffiton and Ammar Malik. Enumerating infeasibility: Finding multiple MUSes quickly. In *CPAIOR*, pages 160–175, 2013. doi:10.1007/978-3-642-38171-3_11.

58 Mark H. Liffiton, Alessandro Previti, Ammar Malik, and Joao Marques-Silva. Fast, flexible MUS enumeration. *Constraints An Int. J.*, 21(2):223–250, 2016. doi:10.1007/S10601-015-9183-0.

59 Xinghan Liu and Emiliano Lorini. A unified logical framework for explanations in classifier systems. *J. Log. Comput.*, 33(2):485–515, 2023. doi:10.1093/LOGCOM/EXAC102.

60 Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, pages 4765–4774, 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

61 Olivier Létoffé, Xuanxiang Huang, and Joao Marques-Silva. Towards rigorous explainability by feature attribution. *CoRR*, abs/2604.15898, April 2026. arXiv:2604.15898.

62 Joao Marques-Silva. Logic-based explainability in machine learning. In *Reasoning Web*, pages 24–104, 2022. doi:10.1007/978-3-031-31414-8_2.

63 Joao Marques-Silva. Disproving XAI myths with formal methods - initial results. In *ICECCS*, pages 12–21, 2023. doi:10.1109/ICECCS59891.2023.00012.

64 Joao Marques-Silva. Logic-based explainability: Past, present and future. In *ISoLA*, pages 181–204, 2024. doi:10.1007/978-3-031-75387-9_12.

65 Joao Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. In *NeurIPS*, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/eccd2a86bae4728b38627162ba297828-Abstract.html.

66 Joao Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explanations for monotonic classifiers. In *ICML*, pages 7469–7479, 2021. URL: http://proceedings.mlr.press/v139/marques-silva21a.html.

67 Joao Marques-Silva, Federico Heras, Mikolás Janota, Alessandro Previti, and Anton Belov. On computing minimal correction subsets. In *IJCAI*, pages 615–622, 2013. URL: http://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6922.

68 Joao Marques-Silva and Xuanxiang Huang. Explainability is *Not* a game. *Commun. ACM*, 67(7):66–75, 2024. doi:10.1145/3635301.

69 Joao Marques-Silva, Xuanxiang Huang, and Olivier Létoffé. The explanation game - rekindled (extended version). *CoRR*, abs/2501.11429, 2025. doi:10.48550/arXiv.2501.11429.

70 Joao Marques-Silva and Alexey Ignatiev. Delivering trustworthy AI through formal XAI. In *AAAI*, pages 12342–12350, 2022. doi:10.1609/AAAI.V36I11.21499.

71 Joao Marques-Silva and Alexey Ignatiev. No silver bullet: interpretable ML models must be explained. *Frontiers Artif. Intell.*, 6, 2023. doi:10.3389/FRAI.2023.1128212.

72 Joao Marques-Silva, Mikolás Janota, and Anton Belov. Minimal sets over monotone predicates in boolean formulae. In *CAV*, pages 592–607, 2013. doi:10.1007/978-3-642-39799-8_39.

73 Joao Marques-Silva, Mikolás Janota, and Carlos Mencía. Minimal sets on propositional formulae. problems and reductions. *Artif. Intell.*, 252:22–50, 2017. doi:10.1016/J.ARTINT.2017.07.005.

74 Joao Marques-Silva, Jairo A. Lefebre-Lobaina, and Maria Vanina Martinez. Efficient and rigorous model-agnostic explanations. In *IJCAI*, pages 2637–2646, 2025. doi:10.24963/IJCAI.2025/294.

75 Joao Marques-Silva and Inês Lynce. On improving MUS extraction algorithms. In *SAT*, pages 159–173, 2011. doi:10.1007/978-3-642-21581-0_14.

76 Hayden McTavish, Zachery Boner, Jon Donnelly, Margo I. Seltzer, and Cynthia Rudin. Leveraging predictive equivalence in decision trees. In *ICML*, 2025. URL: https://proceedings.mlr.press/v267/mctavish25a.html.

77 Carlos Mencía, Raúl Mencía, Ramon Béjar, and Joao Marques-Silva. Model-agnostic explanations by consensus. In *KR*, 2026.

78 Carlos Mencía, Alessandro Previti, and Joao Marques-Silva. Literal-based MCS extraction. In *IJCAI*, pages 1973–1979, 2015. URL: http://ijcai.org/Abstract/15/280.

79 Christoph Molnar. *Interpretable machine learning*. Lulu.com, 2020.

80 Andreas Niskanen, Jeremias Berg, and Matti Järvisalo. Incremental maximum satisfiability. In *SAT*, pages 14:1–14:19, 2022. doi:10.4230/LIPIcs.SAT.2022.14.

81 Alessandro Previti and Joao Marques-Silva. Partial MUS enumeration. In *AAAI*, 2013. doi:10.1609/AAAI.V27I1.8657.

82 Raymond Reiter. A theory of diagnosis from first principles. *Artif. Intell.*, 32(1):57–95, 1987. doi:10.1016/0004-3702(87)90062-2.

83 Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *KDD*, pages 1135–1144, 2016. doi:10.1145/2939672.2939778.

84 Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, pages 1527–1535, 2018. doi:10.1609/AAAI.V32I1.11491.

85 Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi:10.1038/S42256-019-0048-X.

86 Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.

87 Cynthia Rudin and Yaron Shaposhnik. Globally-consistent rule-based summary-explanations for machine learning models: Application to credit-risk evaluation. *J. Mach. Learn. Res.*, 24:16:1–16:44, 2023. URL: https://jmlr.org/papers/v24/21-0488.html.

88 Lloyd S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

89 Lloyd S Shapley and Martin Shubik. A method for evaluating the distribution of power in a committee system. *American political science review*, 48(3):787–792, 1954.

90 Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In *IJCAI*, pages 5103–5111, 2018. doi:10.24963/IJCAI.2018/708.

91 Stephan Wäldchen, Jan MacDonald, Sascha Hauch, and Gitta Kutyniok. The computational complexity of understanding binary classifier decisions. *J. Artif. Intell. Res.*, 70:351–387, 2021. doi:10.1613/JAIR.1.12359.

92 Min Wu, Haoze Wu, and Clark W. Barrett. VeriX: Towards verified explainability of deep neural networks. In *NeurIPS*, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/46907c2ff9fafd618095161d76461842-Abstract-Conference.html.

93 Jinqiang Yu, Graham Farr, Alexey Ignatiev, and Peter J. Stuckey. Anytime approximate formal feature attribution. In *SAT*, pages 30:1–30:23, 2024. doi:10.4230/LIPIcs.SAT.2024.30.

94 Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey, Nina Narodytska, and Joao Marques-Silva. Eliminating the impossible, whatever remains must be true: On extracting and applying background knowledge in the context of formal explanations. In *AAAI*, pages 4123–4131, 2023. doi:10.1609/AAAI.V37I4.25528.