Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Early Career Track

Towards Trustable Explainable AI

Alexey Ignatiev
Monash University, Australia
alexey.ignatiev@monash.edu

Abstract

Explainable artificial intelligence (XAI) represents arguably one of the most crucial challenges being faced by the area of AI these days. Although the majority of approaches to XAI are of heuristic nature, recent work proposed the use of abductive reasoning to computing provably correct explanations for machine learning (ML) predictions. The proposed rigorous approach was shown to be useful not only for computing trustable explanations but also for validating explanations computed heuristically. It was also applied to uncover a close relationship between XAI and verification of ML models. This paper overviews the advances of the rigorous logic-based approach to XAI and argues that it is indispensable if *trustable* XAI is of concern.

1 Introduction

Machine Learning (ML) models are widely used in decision making procedures in many real-world applications. The fast growth, practical achievements and the overall success of modern approaches to ML [LeCun *et al.*, 2015; Jordan and Mitchell, 2015; Mnih *et al.*, 2015; ACM, 2018] guarantees that machine learning will prevail as a generic computing paradigm, and will find an ever growing range of practical applications, many of which will have to do with various aspects of our lives.

Unfortunately, ML models on occasions *catastrophically fail* [Zhou and Sun, 2019; CACM Letters to the Editor, 2019]. They can also support poor decisions due to *bias* (e.g. race, gender, age) in the model [Angwin *et al.*, 2016]. Their decisions can be confusing due to *brittleness* [Szegedy *et al.*, 2014; Goodfellow *et al.*, 2015]. As a result, there is a critical need to understand the behavior of ML models, analyze the (potential) failures of the models (or the data used to train them), debug them, and possibly repair.

This has given rise to a growing interest in validating the operation of ML models [Ruan *et al.*, 2018; Narodytska, 2018; Narodytska *et al.*, 2018b; Katz *et al.*, 2017] but also motivated efforts aiming at devising approaches to explainable artificial intelligence (XAI) [Ribeiro *et al.*, 2018; Lundberg and Lee, 2017; Ignatiev *et al.*, 2018; Narodytska *et al.*, 2018a; Ribeiro *et al.*, 2016; Ignatiev *et al.*, 2019a; Monroe, 2018]. Unfortunately, most existing approaches to explaining ML models are of heuristic nature and are *logically unsound* [Ignatiev *et al.*, 2019c] — one can find counterexamples to heuristic explanations revealing their defects. This exacerbates the problem of trust in AI, as given a misbehaving ML model and an explanation that proves to be incorrect, a user may have even less trust in the ML model.

Recent work [Shih *et al.*, 2018; Ignatiev *et al.*, 2019a] proposed a principled approach to computing *provably correct* explanations (at the expected cost of lower scalability), which is fundamentally different from the heuristic methods. The approach hinges on the use of logic and operates with efficient prime implicant computation for logical representations of the decision function associated with an ML prediction. The approach can be applied directly to the explanation computation but also to validate heuristic explanations [Ignatiev *et al.*, 2019c]. Finally, it has been recently used to establish a rigorous relationship of global explanations and a generalization of adversarial examples, thus, making a bridge between XAI and ML model verification [Ignatiev *et al.*, 2019b]. This paper provides a summary of ongoing efforts to develop formal reasoning approaches for explaining ML models. The reader is referred to the work referenced for further details.

2 Explainable AI and Heuristic Status Quo

Explainable AI is an emerging field, with a growing number of areas of research [Weld and Bansal, 2019]. We focus on explaining predictions of ML models and use the notation of prior work [Shih *et al.*, 2018; Ignatiev *et al.*, 2019a; Ignatiev *et al.*, 2019c; Narodytska *et al.*, 2019; Ignatiev *et al.*, 2019b][1]. Given an instance and its associated prediction, an *explanation* is a set of feature values, with a suitable set of properties. An ML model can be viewed as a function $\mathcal{M} : \mathbb{F} \to \mathbb{K}$, mapping inputs (i.e. the feature values

[1]Here, 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. (The work can be extended to the case of $D_i \subseteq \mathbb{R}$.) An *instance* $\mathcal{I} \in \mathbb{F}$, $\mathbb{F} = D_1 \times \ldots \times D_L$, is a vector of feature values. A classification problem is assumed, with a set of classes $\mathbb{K} = \{\kappa_1, \ldots, \kappa_M\}$. A prediction $\pi \in \mathbb{K}$ is associated with each instance $\mathcal{I} \in \mathbb{F}$. We consider an ML model $\mathbb{M}$, represented by a finite set of first-order logic (FOL) sentences $\mathcal{M}$. (Where viable, alternative representations for $\mathcal{M}$ can be considered, e.g. fragments of FOL, (mixed-)integer linear programming, constraint language(s), etc.)

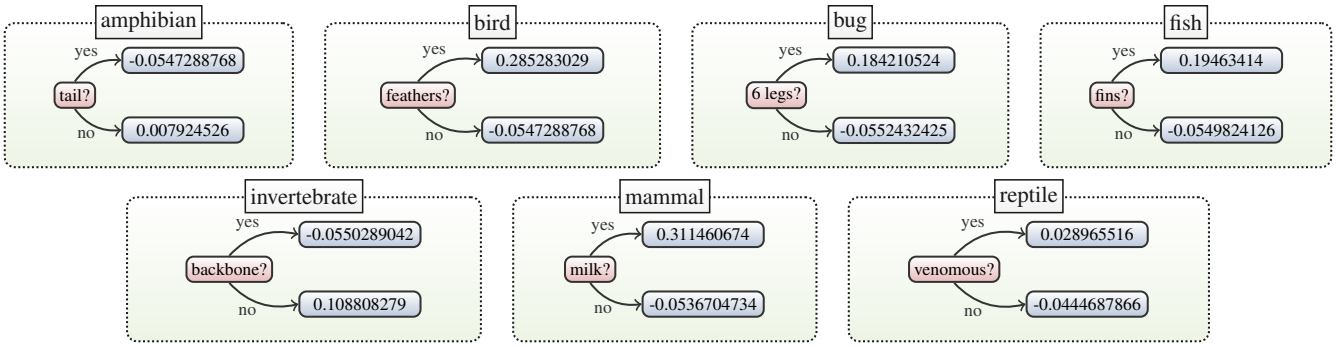

Figure 1: Example of a simplistic boosted tree model [Chen and Guestrin, 2016]. The model targets the well-known Zoo animal classification dataset. Here, the tree ensemble has 1 tree per each of the 7 classes with the depth of each tree being 1.

$\mathbb{F}$) to outputs (i.e. the classes of the classification problem, $\mathbb{K}$). Following [Shih *et al.*, 2018; Ignatiev *et al.*, 2019a], we target computing rigorous explanations of ML predictions. Given some instance $\mathcal{I} \in \mathbb{F}$ and prediction π, a minimal rigorous (or *model-precise*) explanation $\mathcal{E}$ is a prime implicant of the Boolean function $\mathcal{M}(\mathcal{I}) = \pi$. We can also consider minimum explanations, in which case the goal is to compute smallest-size prime implicants.

To illustrate past work on computing explanations, we consider the well-known Zoo dataset[2], generate the boosted tree model shown in Figure 1, and use the following example instance (adapted from [Ignatiev *et al.*, 2019c]):

IF	$(\text{animal\_name} = \text{pitviper}) \land \neg\text{hair} \land \neg\text{feathers} \land$ eggs $\land \neg\text{milk} \land \neg\text{airborne} \land \neg\text{aquatic} \land \text{breathes} \land$ $\neg\text{toothed} \land \text{backbone} \land \text{predator} \land \text{venomous} \land$ $\neg\text{fins} \land (\text{legs} = 0) \land \text{tail} \land \neg\text{domestic} \land \neg\text{catsize}$
THEN	$(\text{class} = \text{reptile})$

The overwhelming majority of work on computing explanations has been on exploiting heuristic approaches, which often offer guarantees of quality with respect to a given instance. Concrete examples include [Baehrens *et al.*, 2010; Ribeiro *et al.*, 2016; Lundberg and Lee, 2017; Ribeiro *et al.*, 2018; Li *et al.*, 2018] among others[3]. LIME [Ribeiro *et al.*, 2016], Anchor [Ribeiro *et al.*, 2018], SHAP [Lundberg and Lee, 2017] represent three alternative approaches for local exploring the feature space with respect to a given instance. As a result, computed explanations are referred to as *local* (or *instance-dependent*).

Despite the prevalence of solutions for computing local explanations heuristically, such explanations are *model-agnostic* and thus offer no guarantees wrt. the underlying ML model nor they provide any guarantees of minimality. For the running example shown earlier, Anchor computes the following explanation:

IF	$\neg\text{hair} \land \neg\text{milk} \land \neg\text{toothed} \land \neg\text{fins}$
THEN	$(\text{class} = \text{reptile})$

[2]https://www.kaggle.com/uciml/zoo-animal-classification.

[3]We focus on the most representative approaches. More comprehensive recent accounts exist [Guidotti *et al.*, 2019].

As Zoo is a very simple dataset, it suffices to analyze the instances in the original dataset to show that the explanation does not hold in general. Concretely, there is at least another instance (in fact, we found a large number of similar examples) for which the Anchor explanation also applies, but for which the boosted tree predicts a *different* class (essentially meaning that the provided explanation is *incorrect*):

IF	$(\text{animal\_name} = \text{toad}) \land \neg\textbf{hair} \land \neg\text{feathers} \land$ eggs $\land \neg\textbf{milk} \land \neg\text{airborne} \land \neg\text{aquatic} \land$ $\neg\text{predator} \land \neg\textbf{toothed} \land \text{backbone} \land \text{breathes} \land$ $\neg\text{venomous} \land \neg\textbf{fins} \land (\text{legs} = 4) \land \neg\text{tail} \land$ $\neg\text{domestic} \land \neg\text{catsize}$
THEN	$(\text{class} = \textbf{amphibian})$

Although one can argue that local explanations are meant to be local to a concrete instance, it is questionable how useful such *heuristic* explanations are for a human decision maker.

3 Trustable Explanations

It is not surprising that heuristic explanations may be incorrect as, being model-agnostic, they are unable to catch all the properties of the underlying ML model. As a result, it is debatable if (and to what extent) heuristic explanations can be trusted. In contrast with heuristic explanations, rigorous logic-based explanations correspond to prime implicants of the Boolean function associated with predicting the class of the target instance, and so hold for *any* point in feature space. In other words, rigorous explanations are provably correct for the entire feature space, which makes them *trustable*.

Two lines of work on computing rigorous explanations have been proposed in recent years. One is based on knowledge compilation [Shih *et al.*, 2018] whilst the other applies abductive reasoning [Ignatiev *et al.*, 2019a] detailed below. Knowledge compilation aims at finding canonical representations of functions, which in turn enable efficient algorithms for answering queries that would be too inefficient on the original formula. Although effective once the target representation is obtained, the compilation process itself is worst-case exponential in time and space, which may represent a significant obstacle. Also, the use of compilation for computing explanations requires developing dedicated algorithms for each ML model.

3.1 Trustable Explanations with Abduction

Rigorous explanations can be obtained by abductive reasoning [Ignatiev *et al.*, 2016; Ignatiev *et al.*, 2019a], without the explicit need of computing a canonical representation of the function associated with the ML model. An advantage of this approach is that the common limitations regarding the size of the representation do not apply. A drawback is that an explanation needs to be computed for each instance, whereas with compilation-based approaches explanations for any instance become available as soon as the target representation is obtained. In practice, the use of abductive reasoning requires a logic-based representation to be devised for a given ML model. Earlier work considered neural networks [Ignatiev *et al.*, 2019a] and boosted trees [Ignatiev *et al.*, 2019c].

Given a dataset, we can use any existing approach for computing an ML model. For the boosted tree of Figure 1, one can devise an encoding using ILP or SMT [Ignatiev *et al.*, 2019c]. Given the running example, we can compute a prime implicant which, given the model subject to the prediction, entails the prediction. In this case, using abductive reasoning [Ignatiev *et al.*, 2019a; Ignatiev *et al.*, 2019c], a possible explanation is:

IF	$\neg$feathers $\land$ $\neg$milk $\land$ backbone $\land$ $\neg$fins $\land$ (legs $=0$) $\land$ tail
THEN	(class $=$ reptile)

This *rigorous* explanation guarantees that, as long as the six indicated literals take the value shown, the prediction will be the same, *independently* of the value of *any* of the other features. More importantly, the use of SMT or SAT reasoners enables the generation of a *proof* (or *proof trace*) for each computed explanation, which can be independently validated. This offers another degree of trust in computed explanations. Furthermore, there is no conceptual obstacle to extracting proofs or proof traces from other types of reasoners, e.g. ILP.

Recent work [Ignatiev *et al.*, 2019a] has shown that the use of abductive reasoning for computing explanations can analyze neural networks of modest size. In contrast, for boosted trees, abductive reasoning can be applied to trees of the sizes of practical interest [Ignatiev *et al.*, 2019c].

4 Assessing and Repairing Heuristic Methods

The observations above raise concerns regarding the quality of heuristic explanations computed by approaches such as LIME [Ribeiro *et al.*, 2016], SHAP [Lundberg and Lee, 2017], or Anchor [Ribeiro *et al.*, 2018]. This section outlines recent results on assessing the quality of heuristic explanations. One validates them by abductive reasoning [Ignatiev *et al.*, 2019c], whereas the other exploits recent work on approximate model counting [Narodytska *et al.*, 2019].

4.1 Validating Heuristic Explanations

By applying formal reasoning about the logic-based representation of a classifier, one can check whether or not an explanation holds in the *entire instance space*, i.e. there exist counterexamples [Ignatiev *et al.*, 2019c]. Such reasoning, thus, represents a principled approach for assessing validity of heuristic explanations. Furthermore, in case the explanation is invalid, one can make an attempt to repair it using a series of reasoning oracles calls. Also, as heuristic approaches like LIME, Anchor, and SHAP do not guarantee minimality of the explanation produced, one can try to reduce it further in case it is valid. A possible setup for such explanation validation procedure is shown in Figure 2. Given an explanation, its validity is checked in the entire instance space. If the explanation is invalid, it is reported to be *incorrect* and can be repaired. Otherwise, the explanation is proved to be valid and an attempt to reduce it is made. If the attempt succeeds, the explanation is deemed *redundant*. Otherwise, the explanation is proved to be correct and minimal.

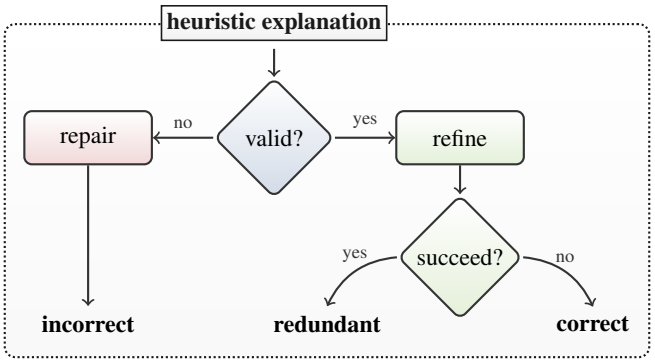

Figure 2: Assessing the validity of a heuristic explanation.

Table 1 summarizes the experiment detailed in [Ignatiev *et al.*, 2019c], which given a trained boosted tree computes a heuristic explanation for *each unique* instance of an input dataset, either with LIME, Anchor, or SHAP and assesses the explanation's correctness following the setup of Figure 2. Five publicly available datasets are considered in the experiment. Three of them were studied in [Ribeiro *et al.*, 2018] to illustrate the advantage of Anchor over LIME, including *adult*, *lending*, and *recidivism*. Two more datasets (*compas* and *german*) were previously studied in the context of algorithmic fairness (e.g. see [Angwin *et al.*, 2016]).

As can be observed, most explanations computed by LIME, Anchor, and SHAP are inadequate, from the global perspective. Observe that for the 4 out of 5 datasets the explanations of all three explainers are mostly incorrect. As an example, for *recidivism* and *german* more than 99% of Anchor's explanations are invalid. Similar results hold for LIME (SHAP, resp.), i.e. 94.1% and 85.3% (85.9% and 63.0%) explanations for *recidivism* and *german* are incorrect. Also note that the number of redundant explanations is usually lower, with the exception of SHAP. Overall and with the exception of *lending*, the number of correct explanations does not go beyond 17.9% for Anchor, 30.8% for LIME, and 19.1% for SHAP and usually constitutes just a few percent.

4.2 Evaluating Quality of Heuristic Explanations

Earlier work [Narodytska *et al.*, 2019] proposed a generic strategy to assess the quality of heuristic explanations using approximate model counting [Soos and Meel, 2019]. Here, we overview the quality assessment results for Anchor

Dataset	(# unique)	Explanations								
		incorrect			redundant			correct		
		LIME	Anchor	SHAP	LIME	Anchor	SHAP	LIME	Anchor	SHAP
adult	(5579)	61.3%	80.5%	70.7%	7.9%	1.6%	10.2%	30.8%	17.9%	19.1%
lending	(4414)	24.0%	3.0%	17.0%	0.4%	0.0%	2.5%	75.6%	97.0%	80.5%
rcdv	(3696)	94.1%	99.4%	85.9%	4.6%	0.4%	7.9%	1.3%	0.2%	6.2%
compas	(778)	71.9%	84.4%	60.4%	20.6%	1.7%	27.8%	7.5%	13.9%	11.8%
german	(1000)	85.3%	99.7%	63.0%	14.6%	0.2%	37.0%	0.1%	0.1%	0.0%

Table 1: Heuristic explanations assessed, for each data instance of the input datasets. The table shows the percentage of incorrect, redundant, and correct explanations provided by LIME, Anchor, and SHAP. The total number of unique instances per dataset is shown in column 2.

Dataset	Unconstrained inputs		Constrainted inputs	
	Anchor	ApproxMC3	Anchor	ApproxMC3
adult	0.99	0.67	0.99	0.81
lending	0.99	0.87	0.99	0.92
rcdv	0.99	0.75	0.99	0.80

Table 2: The precision metric estimates by ApproxMC3 and Anchor (the average over 300 samples).

applied to binarized neural networks [Hubara *et al.*, 2016; Narodytska *et al.*, 2018b]. We considered the three above datasets: *adult*, *lending*, and *recidivism*. Anchor's *precision estimate* used is defined in [Ribeiro *et al.*, 2018], as a quality metric. To perform model counting, we used the approximate model counting solver ApproxMC3 with the standard tolerance and confidence ($\varepsilon = 0.8$ and $\delta = 0.2$) [Soos and Meel, 2019]. Two sets of experiments were performed: (1) *constrained* and (2) *unconstrained* — depending on whether or not the instance space was restricted to be a local neighborhood of a given instance. We worked with instances for which Anchor's precision estimates were around 0.99. Table 2 summarizes our results. Note that the quality of Anchor's explanations can vary wildly, indicating that for some datasets Anchor's explanations are fairly accurate, but in some other cases they can be quite inaccurate. For example, observe that Anchor's estimates of the precision metric are good for the *lending* dataset. On average, the discrepancy between Anchor's estimates and our assessment was 0.1 for this set. In contrast, the average discrepancy was high in the *adult* dataset, 0.25. Note that in our experiments, ApproxMC3 had a significant theoretical error bound on the estimate we produce. Namely, we could be up to 80% off the true solution count in the worst case. Theoretically, our framework allows us to obtain an estimate with much tighter theoretical guarantees but computing these estimates is computationally expensive for the benchmarks considered. However, studies have reported that the tolerance observed in the experiments is far better than the theoretical guarantee [Soos and Meel, 2019].

The negative results summarized in this section assume that, in order to be trustable, explanations are to be meaningful over the entire feature space. This is what rigorous explanations offer. If for some reason the entire feature space is not relevant, the computation of rigorous explanations can also take this information into account.

5 Relating XAI and ML Model Verification

Adversarial examples (AE's) [Goodfellow *et al.*, 2015] illustrate the brittleness of machine learning (ML) models, and have been the subject of growing interest in recent years. Over the last few years, a number of works realized the existence of some connection between AE's and XP's [Tao *et al.*, 2018; Chalasani *et al.*, 2018]. Recent work [Ignatiev *et al.*, 2019b] exploited *global* rigorous explanations, introduced the concept of *counterexamples*, tightly related with adversarial examples, and then showed a minimal hitting set duality relationship between explanations and counterexamples, and in the process proposed approaches for computing adversarial examples from explanations and vice-versa. These recent results build on work from the 80s and 90s, first on model-based diagnosis [Reiter, 1987], and then on computing primes of (Boolean) functions [Rymon, 1994], but also suggest that additional connections between analysis of ML models and other areas of AI can be established.

6 Conclusions

This paper overviews recent work on exploiting the logic-based approach for computing rigorous explanations to ML predictions. Moreover, the paper argues that rigorous explanations are *trustable*, in clear contrast to heuristic explanations, which is confirmed by the experimental results casting doubt on the quality of the heuristic explanations of [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017; Ribeiro *et al.*, 2018]. We conjecture that potential incorrectness is intrinsic to approaches that compute model-agnostic heuristic explanations. Finally, the paper summarizes recent results relating global rigorous explanations with adversarial examples. These results build on earlier seminal work on relating conflicts and diagnoses in model-based diagnosis through subset-minimal hitting sets [Reiter, 1987], but also on finding prime implicants and implicates [Rymon, 1994].

Acknowledgements

The author thanks his colleagues Joao Marques-Silva and Nina Narodytska, who have been taking active part in the research on rigorous logic-based XAI and coauthoring the works, which this paper extensively builds on. Without them this work would be impossible.

References

[ACM, 2018] ACM. Fathers of the deep learning revolution receive ACM A.M. Turing award. http://tiny.cc/9plzpz, 2018.

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. http://tiny.cc/dd7mjz, 2016.

[Baehrens *et al.*, 2010] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831, 2010.

[CACM Letters to the Editor, 2019] CACM Letters to the Editor. A case against mission-critical applications of machine learning. *Commun. ACM*, 62(8):9, 2019.

[Chalasani *et al.*, 2018] Prasad Chalasani, Somesh Jha, Aravind Sadagopan, and Xi Wu. Adversarial learning and explainability in structured datasets. *CoRR*, abs/1810.06583, 2018.

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

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR (Poster)*, 2015.

[Guidotti *et al.*, 2019] 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.

[Hubara *et al.*, 2016] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *NIPS*, pages 4107–4115, 2016.

[Ignatiev *et al.*, 2016] Alexey Ignatiev, Antonio Morgado, and Joao Marques-Silva. Propositional abduction with implicit hitting sets. In *ECAI*, pages 1327–1335, 2016.

[Ignatiev *et al.*, 2018] Alexey Ignatiev, Filipe Pereira, Nina Narodytska, and Joao Marques-Silva. A SAT-based approach to learn explainable decision sets. In *IJCAR*, pages 627–645, 2018.

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

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

[Ignatiev *et al.*, 2019c] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On validating, repairing and refining heuristic ML explanations. *CoRR*, abs/1907.02509, 2019.

[Jordan and Mitchell, 2015] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[Katz *et al.*, 2017] 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.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[Li *et al.*, 2018] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *AAAI*, pages 3530–3537, 2018.

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

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[Monroe, 2018] Don Monroe. AI, explain yourself. *Commun. ACM*, 61(11):11–13, 2018.

[Narodytska *et al.*, 2018a] Nina Narodytska, Alexey Ignatiev, Filipe Pereira, and Joao Marques-Silva. Learning optimal decision trees with SAT. In *IJCAI*, pages 1362–1368, 2018.

[Narodytska *et al.*, 2018b] Nina Narodytska, Shiva Prasad Kasiviswanathan, Leonid Ryzhyk, Mooly Sagiv, and Toby Walsh. Verifying properties of binarized deep neural networks. In *AAAI*, 2018.

[Narodytska *et al.*, 2019] Nina Narodytska, Aditya A. Shrotri, Kuldeep S. Meel, Alexey Ignatiev, and Joao Marques-Silva. Assessing heuristic machine learning explanations with model counting. In *SAT*, pages 267–278, 2019.

[Narodytska, 2018] Nina Narodytska. Formal analysis of deep binarized neural networks. In *IJCAI*, pages 5692–5696, 2018.

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

[Ribeiro *et al.*, 2016] 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.

[Ribeiro *et al.*, 2018] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, 2018.

[Ruan *et al.*, 2018] Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. Reachability analysis of deep neural networks with provable guarantees. In *IJCAI*, pages 2651–2659, 2018.

[Rymon, 1994] Ron Rymon. An SE-tree-based prime implicant generation algorithm. *Ann. Math. Artif. Intell.*, 11(1-4):351–366, 1994.

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

[Soos and Meel, 2019] Mate Soos and Kuldeep S. Meel. BIRD: Engineering an efficient CNF-XOR SAT solver and its applications to approximate model counting. In *AAAI*, 2019.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR (Poster)*, 2014.

[Tao *et al.*, 2018] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In *NeurIPS*, pages 7728–7739, 2018.

[Weld and Bansal, 2019] Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Commun. ACM*, 62(6):70–79, 2019.

[Zhou and Sun, 2019] Zhi Quan Zhou and Liqun Sun. Metamorphic testing of driverless cars. *Commun. ACM*, 62(3):61–67, 2019.