Explainability is NOT a Game

Joao Marques-Silva, Xuanxiang Huang

Explainability is NOT a Game

Joao Marques-Silva*
joao.marques-silva@irit.fr
IRIT, CNRS
Toulouse, France

Xuanxiang Huang*
xuanxiang.huang@univ-toulouse.fr
University of Toulouse
Toulouse, France

Abstract

Explainable artificial intelligence (XAI) aims to help human decision-makers in understanding complex machine learning (ML) models. One of the hallmarks of XAI are measures of relative feature importance, which are theoretically justified through the use of Shapley values. This paper builds on recent work and offers a simple argument for why Shapley values can provide misleading measures of relative feature importance, by assigning more importance to features that are irrelevant for a prediction, and assigning less importance to features that are relevant for a prediction. The significance of these results is that they effectively challenge the many proposed uses of measures of relative feature importance in a fast-growing range of high-stakes application domains.

CCS Concepts

• Computing methodologies → Artificial intelligence; Machine learning algorithms; Machine learning; • Theory of computation → Automated reasoning.

Keywords

Explainable AI, Shapley values, Abductive reasoning

1 Introduction

The societal and economic significance of machine learning (ML) cannot be overstated, with many remarkable advances made in recent years. However, the operation of complex ML models is most often inscrutable, with the consequence that decisions taken by ML models cannot be fathomed by human decision makers. It is therefore of importance to devise automated approaches to explain the predictions made by complex ML models. This is the main motivation for eXplainable AI (XAI). Explanations thus serve to build trust, but also to debug complex systems of AI. Furthermore, in situations where decisions of ML models impact people, one should expect explanations to offer the strongest guarantees of rigor.

However, the most popular XAI approaches [Bach et al. 2015; Lundberg and Lee 2017; Ribeiro et al. 2016, 2018; Samek et al. 2019] offer no guarantees of rigor. Unsurprisingly, a number of works have demonstrated several misconceptions of informal approaches to XAI [Huang and Marques-Silva 2023; Ignatiev 2020; Ignatiev et al. 2019b; Izza et al. 2022; Marques-Silva 2023; Narodytska et al. 2019]. In contrast to informal XAI, formal explainability offers a logic-based, model-precise approach for computing explanations [Ignatiev et al. 2019a]. Although formal explainability also exhibits a number of drawbacks, including the computational complexity of logic-based reasoning, there has been continued progress since its inception [Marques-Silva 2022; Marques-Silva and Ignatiev 2022].

Among the existing informal approaches to XAI, the use of Shapley values as a mechanism for feature attribution is arguably the best-known. Shapley values [Shapley 1953] were originally proposed in the context of game theory, but have found a wealth of application domains [Roth 1988]. More importantly, for more than two decades Shapley values have been proposed in the context of explaining the decisions of complex ML models [Lipovetsky and Conklin 2001; Lundberg and Lee 2017; Strumbelj and Kononenko 2010, 2014]. The importance of Shapley values for explainability is illustrated by the massive impact of tools like SHAP [Lundberg and Lee 2017], including many recent uses that have a direct influence on human beings (see [Huang and Marques-Silva 2023] for some recent references).

Unfortunately, the exact computation of Shapley values in the case of explainability has not been studied in practice, in part because of its computational complexity. Hence, it is unclear how good are existing approximate solutions, with a well-known example being SHAP [Chen et al. 2022; Lundberg et al. 2020; Lundberg and Lee 2017]. Recent work [Arenas et al. 2021b] proposed a polynomial-time algorithm for computing Shapley values in the case of classifiers represented by deterministic decomposable boolean circuits. As a result, and for one concrete family of classifiers, it became possible to compare the estimates of tools such as SHAP [Lundberg and Lee 2017] with those obtained with exact algorithms.

Furthermore, since Shapley values aim to measure the relative importance of features, a natural question is whether the relative importance of features obtained with Shapley values can indeed be trusted. Given that the definition of Shapley values is axiomatic, one may naturally question how reliable those values are. Evidently, if the relative order of features dictated by Shapley values can be proved inadequate, then the use of Shapley values in explainability ought to be deemed unworthy of trust.

A number of earlier works reported practical problems with explainability approaches based on Shapley values [Kumar et al. 2020] ([Huang and Marques-Silva 2023] covers a number of additional references). However, these works focus on practical tools, which approximate Shapley values,

*Both authors contributed equally to this research.

but do not investigate the possible existence of fundamental limitations with the use of Shapley values in explainability. In contrast with these other works, this paper offers a simple argument for why relative feature importance obtained with Shapley values can provide misleading information, in that features that bear no significance for a prediction can be deemed more important, in terms of Shapley values, than features that bear some significance for the same prediction. The importance of this paper's results, and of the identified flaws with Shapley values, should be assessed in light of the fast-growing uses of explainability solutions in domains that directly impact human beings, e.g. medical diagnostic applications, especially when the vast majority of such uses build on Shapley values for explainability.

The paper is organized as follows. Section 2 introduces the notation and definitions used throughout. This includes a brief introduction to formal explanations, but also to Shapley values for explainability. Section 3 revisits the concepts of relevancy/irrelevancy, which have been studied in logic-based abduction since the mid 1990s [Eiter and Gottlob 1995]. Section 4 demonstrates the inadequacy of Shapley values for feature attribution. Finally, Section 5 discusses the paper's results, but it also briefly examines additional flaws of Shapley values.

2 Definitions

Throughout the paper, we adopt the notation and the definitions introduced in earlier work, namely [Marques-Silva 2022; Marques-Silva and Ignatiev 2022] and also [Arenas et al. 2021b].

2.1 Classification Problems

A classification problem is defined on a set of features $\mathcal{F} = \{1, \ldots, m\}$, and a set of classes $\mathcal{K} = \{c_1, \ldots, c_K\}$. Each feature $i \in \mathcal{F}$ takes values from a domain $\mathcal{D}_i$. Domains can be ordinal (e.g. real- or integer-valued) or categorical. Feature space is defined by the cartesian product of the domains of the features: $\mathbb{F} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_m$. A classifier $\mathcal{M}$ computes a (non-constant) classification function: $\kappa : \mathbb{F} \to \mathcal{K}$[1]. A classifier $\mathcal{M}$ is associated with a tuple $(\mathcal{F}, \mathbb{F}, \mathcal{K}, \kappa)$. For the purposes of this paper, we restrict κ to be a non-constant boolean function. This restriction does not in any way impact the validity of our results.

Given a classifier $\mathcal{M}$, and a point $\mathbf{v} \in \mathbb{F}$, with $c = \kappa(\mathbf{v})$ and $c \in \mathcal{K}$, $(\mathbf{v}, c)$ is referred to as an instance (or sample). An explanation problem $\mathcal{E}$ is associated with a tuple $(\mathcal{M}, (\mathbf{v}, c))$. As a result, $\mathbf{v}$ represents a concrete point in feature space, whereas $\mathbf{x} \in \mathbb{F}$ represents an arbitrary point in feature space.

As a running example, we consider the decision tree (DT) shown in Figure 1. Since it will be used later, we also show the truth table for the DT classifier. Given the information shown in the DT, we have that $\mathcal{F} = \{1, 2, 3, 4\}$, $\mathcal{D}_i = \{0, 1\}$, $i = 1, 2, 3, 4$, $\mathbb{F} = \{0, 1\}^4$, and $\mathcal{K} = \{0, 1\}$. The classification function κ is given by the decision tree shown, or alternatively by the truth table. Finally, the instance considered is $(\mathbf{v}, c) = ((0, 0, 0, 0), 0)$.

2.2 Formal Explanations

The presentation of formal explanations follows recent accounts [Marques-Silva 2022]. In the context of XAI, abductive explanations (AXp's) have been studied since 2018 [Ignatiev et al. 2019a; Shih et al. 2018]. Similar to other heuristic approaches, e.g. Anchors [Ribeiro et al. 2018], abductive explanations are an example of explainability by feature selection, i.e. a subset of features is selected as the explanation. AXp's represent a rigorous example of explainability by feature selection, and can be viewed as the answer to a "Why (the prediction)?" question. An AXp is defined as a subset-minimal (or irreducible) set of features $\mathcal{X} \subseteq \mathcal{F}$ such that the features in $\mathcal{X}$ are sufficient for the prediction. This is to say that, if the features in $\mathcal{X}$ are fixed to the values determined by $\mathbf{v}$, then the prediction is guaranteed to be $c = \kappa(\mathbf{v})$. The sufficiency for the prediction can be stated formally:

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

Observe that (1) is monotone on $\mathcal{X}$, and so the two conditions for a set $\mathcal{X} \subseteq \mathcal{F}$ to be an AXp (i.e. sufficiency for prediction and subset-minimality), can be stated as follows:

$$\begin{aligned} &\forall(\mathbf{x} \in \mathbb{F}).\ \left[\bigwedge\nolimits_{i \in \mathcal{X}} (x_i = v_i)\right] \rightarrow (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \wedge \\ &\forall(t \in \mathcal{X}).\exists(\mathbf{x} \in \mathbb{F}).\ \left[\bigwedge\nolimits_{i \in \mathcal{X} \setminus \{t\}} (x_i = v_i)\right] \wedge (\kappa(\mathbf{x}) \neq \kappa(\mathbf{v})) \end{aligned} \tag{2}$$

A predicate $\mathsf{AXp} : 2^{\mathcal{F}} \to \{0, 1\}$ is associated with (2), such that $\mathsf{AXp}(\mathcal{X}; \mathcal{E})$ holds true if and only if (2) holds true[2].

An AXp can be interpreted as a logic rule of the form:

$$\text{IF} \quad \left[\bigwedge\nolimits_{i \in \mathcal{X}} (x_i = v_i)\right] \quad \text{THEN} \quad (\kappa(\mathbf{x}) = c) \tag{3}$$

where $c = \kappa(\mathbf{v})$. It should be noted that informal XAI methods have also proposed the use of IF-THEN rules [Ribeiro et al. 2018] which, in the case of Anchors [Ribeiro et al. 2018] may or may not be sound [Ignatiev 2020; Ignatiev et al. 2019a]. In contrast, rules obtained from AXp's are logically sound.

Moreover, contrastive explanations (CXp's) represent a type of explanation that differs from AXp's, in that CXp's answer a "Why Not (some other prediction)?" question [Ignatiev et al. 2020; Miller 2019]. Given a set $\mathcal{Y} \subseteq \mathcal{F}$, sufficiency for changing the prediction can be stated formally:

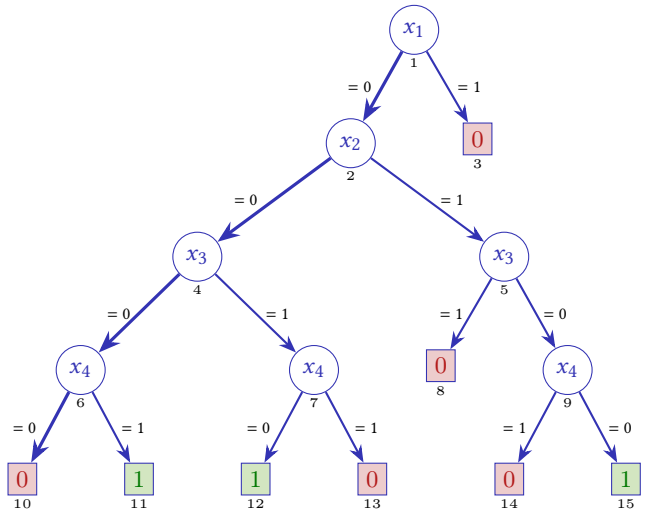
$$\exists(\mathbf{x} \in \mathbb{F}).\ \left[\bigwedge\nolimits_{i \in \mathcal{F} \setminus \mathcal{Y}} (x_i = v_i)\right] \wedge (\kappa(\mathbf{x}) \neq \kappa(\mathbf{v})) \tag{4}$$

A CXp is a subset-minimal set of features which, if allowed to take a value other than the value determined by $\mathbf{v}$, then the predicted can be changed by choosing suitable values to those features.

Similarly to the case of AXp's, for CXp's (4) is monotone on $\mathcal{Y}$, and so the two conditions (sufficiency for changing the

[1] A classifier that computes a constant function, i.e. the same prediction for all points in feature space, is of course uninteresting, and so it is explicitly disallowed.

[2] When defining concepts, we will show the necessary parameterizations. However, in later uses, those parameterizations will be omitted, for simplicity.



row #	x_1	x_2	x_3	x_4	$\kappa(\mathbf{x})$
1	0	0	0	0	0
2	0	0	0	1	1
3	0	0	1	0	1
4	0	0	1	1	0
5	0	1	0	0	1
6	0	1	0	1	0
7	0	1	1	0	0
8	0	1	1	1	0
9	1	0	0	0	0
10	1	0	0	1	0
11	1	0	1	0	0
12	1	0	1	1	0
13	1	1	0	0	0
14	1	1	0	1	0
15	1	1	1	0	0
16	1	1	1	1	0

Figure 1: Example classifier – decision tree and its truth table. For this classifier, we have $\mathcal{F} = \{1,2,3,4\}$, $\mathcal{D}_i = \{0,1\}, i = 1,2,3,4$, $\mathbb{F} = \{0,1\}^4$, and $\mathcal{K} = \{0,1\}$. The classification function is given by the decision tree shown, or alternatively by the truth table. Finally, the instance considered is $((0,0,0,0),0)$, corresponding to row 1 in the truth table. The instance is consistent with path $\langle 1,2,4,6,10\rangle$, which is highlighted in the DT. The prediction is 0, as indicated in terminal node 10.

prediction and subset-minimality) can be stated formally as follows:

$$\exists(\mathbf{x}\in\mathbb{F}).\left[\bigwedge\nolimits_{i\in\mathcal{F}\setminus\mathcal{Y}}(x_i=v_i)\right]\land(\kappa(\mathbf{x})\neq\kappa(\mathbf{v}))\land \tag{5}$$
$$\forall(t\in\mathcal{Y}).\forall(\mathbf{x}\in\mathbb{F}).\left[\bigwedge\nolimits_{i\in\mathcal{F}\setminus(\mathcal{Y}\setminus\{t\})}(x_i=v_i)\right]\rightarrow(\kappa(\mathbf{x})=\kappa(\mathbf{v}))$$

A predicate $\mathsf{CXp}:2^{\mathcal{F}}\rightarrow\{0,1\}$ is associated with (5), such that $\mathsf{CXp}(\mathcal{Y};\mathcal{E})$ holds true if and only if (5) holds true.

Algorithms for computing AXp's and CXp's for different families of classifiers have been proposed in recent years ([Marques-Silva and Ignatiev 2022] provides a recent account of the progress observed in computing formal explanations). These algorithms include the use of automated reasoners (e.g. SAT, SMT or MILP solvers), or dedicated algorithms for families of classifiers for which computing one explanation is tractable.

Given an explanation problem $\mathcal{E}$, the sets of AXp's and CXp's are represented by:

$$\mathbb{A}(\mathcal{E})=\{\mathcal{X}\subseteq\mathcal{F}\,|\,\mathsf{AXp}(\mathcal{X};\mathcal{E})\} \tag{6}$$
$$\mathbb{C}(\mathcal{E})=\{\mathcal{Y}\subseteq\mathcal{F}\,|\,\mathsf{CXp}(\mathcal{Y};\mathcal{E})\} \tag{7}$$

For example, $\mathbb{A}(\mathcal{E})$ represents the set of all logic rules that predict $c=\kappa(\mathbf{v})$, which are consistent with $\mathbf{v}$, and which are irreducible (i.e. no literal $x_i=v_i$ can be discarded).

Furthermore, it has been proved [Ignatiev et al. 2020] that (i) a set $\mathcal{X}\subseteq\mathcal{F}$ is an AXp if and only if it is a minimal hitting set (MHS) of the set of CXp's; and (ii) a set $\mathcal{Y}\subseteq\mathcal{F}$ is a CXp if and only if it is an MHS of the set of AXp's. This property is referred to as MHS duality, and can be traced back to the seminal work of R. Reiter [Reiter 1987] in model-based diagnosis. Moreover, MHS duality has been shown to be instrumental for the enumeration of AXp's and CXp's,

$\mathcal{S}$	Template	Rows	Points	Value	$\kappa(\mathbf{x})=c$?
$\{2,3\}$	$(x_1,0,0,x_4)$	1	$(0,0,0,0)$	0	No
		2	$(0,0,0,1)$	1	
		9	$(1,0,0,0)$	0	
		10	$(1,0,0,1)$	0	
$\{1,2,4\}$	$(0,0,x_3,0)$	1	$(0,0,0,0)$	0	No
		3	$(0,0,1,0)$	1	
$\{2,3,4\}$	$(x_1,0,0,0)$	1	$(0,0,0,0)$	0	Yes
		9	$(1,0,0,0)$	0	

Table 1: Examples of how each set is analyzed when computing AXp's. For CXp's, a similar approach is used.

but also for answering other explainability queries [Marques-Silva 2022].

For the running example, and since it is feasible to represent the function with a truth table, then there exist polynomial-time algorithms (on the size of the truth-table) for computing all AXp's and all CXp's [Huang and Marques-Silva 2023]. This is illustrated in Figure 2. Table 1 illustrates how each set is analyzed when computing AXp's or CXp's.

Formal explainability has made significant progress in recent years, covering a wide range of topics of research. [Marques-Silva 2022] represents a recent overview of the progress in the emerging field of formal explainability.

2.3 Shapley Values in Explainability

Shapley values were proposed in the 1950s, in the context of game theory [Shapley 1953], and find a wealth of uses [Roth 1988]. More recently, Shapley values have been

$\mathcal{S}$	rows picked by $\mathcal{S}$	$\kappa(\mathbf{x}) = c$?	$\mathcal{S}$ is AXp?	$\mathcal{F} \setminus \mathcal{S}$	rows picked by $\mathcal{F} \setminus \mathcal{S}$	$\kappa(\mathbf{x}) \neq c$?	$\mathcal{S}$ is CXp?
$\emptyset$	1..16	No		$\{1,2,3,4\}$	1	No	
$\{1\}$	1,2,3,4,5,6,7,8	No		$\{2,3,4\}$	1,9	No	
$\{2\}$	1,2,3,4,9,10,11,12	No		$\{1,3,4\}$	1,5	Yes	Yes
$\{3\}$	1,2,5,6,9,10,13,14	No		$\{1,2,4\}$	1,3	Yes	Yes
$\{4\}$	1,3,5,7,9,11,13,15	No		$\{1,2,3\}$	1,2	Yes	Yes
$\{1,2\}$	1,2,3,4	No		$\{3,4\}$	1,5,9,13	Yes	
$\{1,3\}$	1,2,5,6	No		$\{2,4\}$	1,3,9,11	Yes	
$\{1,4\}$	1,3,5,7	No		$\{2,3\}$	1,2,9,10	Yes	
$\{2,3\}$	1,2,9,10	No		$\{1,4\}$	1,3,5,7	Yes	
$\{2,4\}$	1,3,9,11	No		$\{1,3\}$	1,2,5,6	Yes	
$\{3,4\}$	1,5,9,13	No		$\{1,2\}$	1,2,3,4	Yes	
$\{1,2,3\}$	1,2	No		$\{4\}$	1,3,5,7,9,11,13,15	Yes	
$\{1,2,4\}$	1,3	No		$\{3\}$	1,2,5,6,9,10,13,14	Yes	
$\{1,3,4\}$	1,5	No		$\{2\}$	1,2,3,4,9,10,11,12	Yes	
$\{2,3,4\}$	1,9	Yes	Yes	$\{1\}$	1,2,3,4,5,6,7,8	Yes	
$\{1,2,3,4\}$	1	Yes		$\emptyset$	1..16	Yes	

Figure 2: Computing AXp's/CXp's for the example DT and instance $((0,0,0,0),0)$. All subsets of features are considered. For computing AXp's, and for some set $\mathcal{S}$, the features in $\mathcal{S}$ are fixed to their values as dictated by $\mathbf{v}$. The picked rows are the rows consistent with those fixed values. For example, if $\mathcal{S} = \{2,3,4\}$, then only rows 1 and 9 are consistent with having features 2, 3 and 4 assigned value 0. Similarly, for computing CXp's, and for some set $\mathcal{S}$, the features in $\mathcal{F} \setminus \mathcal{S}$ are fixed to their values as dictated by $\mathbf{v}$. The picked rows are again the rows consistent with those fixed values. For example, if $\mathcal{S} = \{2\}$, then $\mathcal{F} \setminus \mathcal{S} = \{1,3,4\}$, and so only rows 1 and 5 are consistent with having features 1, 3 and 4 assigned value 0. An AXp is an irreducible set of features that is sufficient for the prediction. In this example, only $\{2,3,4\}$ respects the criteria. Moreover, a CXp is an irreducible set of features which, if allowed to take any value from their domain, the prediction changes. For this example, $\{2\}$, $\{3\}$ and $\{4\}$ respect the criteria, i.e. by only changing one of these features, we are able to change the prediction.

extensively used for explaining the predictions of ML models, e.g. [Chen et al. 2019; Datta et al. 2016; Lipovetsky and Conklin 2001; Lundberg and Lee 2017; Merrick and Taly 2020; Slack et al. 2021; Strumbelj and Kononenko 2010, 2014; Watson 2022], among a vast number of recent examples (see [Huang and Marques-Silva 2023] for a more comprehensive list of references). Shapley values represent one example of explainability by feature attribution, i.e. some score is assigned to each feature as a form of explanation. The complexity of computing Shapley values (as proposed in SHAP [Lundberg and Lee 2017]) has been studied in recent years [Arenas et al. 2021a,b; den Broeck et al. 2021, 2022]. This section provides a brief overview of how Shapley values for explainability are computed. Throughout, we build on the notation used in recent work [Arenas et al. 2021a,b], which builds on the work of [Lundberg and Lee 2017].

Let $\Upsilon : 2^{\mathcal{F}} \to 2^{\mathbb{F}}$ be defined by,

$$\Upsilon(\mathcal{S};\mathbf{v}) = \{\mathbf{x} \in \mathbb{F} \,|\, \wedge_{i \in \mathcal{S}}\, x_i = v_i\} \tag{8}$$

i.e. for a given set $\mathcal{S}$ of features, and parameterized by the point $\mathbf{v}$ in feature space, $\Upsilon(\mathcal{S};\mathbf{v})$ denotes all the points in feature space that have in common with $\mathbf{v}$ the values of the features specified by $\mathcal{S}$. Observe that Υ is also used (implicitly) for picking the set of rows where are interested in when computing explanations (see Table 1).

$\mathcal{S}$	Template	$\Upsilon(\mathcal{S};\mathbf{v})$	Rows	$\phi(\mathcal{S})$
$\{1,4\}$	$(0,x_2,x_3,0)$	$\{(0,0,0,0),$ $(0,0,1,0),$ $(0,1,0,0),$ $(0,1,1,0)\}$	$\{1,3,5,7\}$	$2/4$
$\{3,4\}$	$(x_1,x_2,0,0)$	$\{(0,0,0,0),$ $(0,1,0,0),$ $(1,0,0,0),$ $(1,1,0,0)\}$	$\{1,5,9,13\}$	$1/4$

Table 2: Computation of average values. Rows represents the numbers of rows to consider when computing the average value.

Also, let $\phi : 2^{\mathcal{F}} \to \mathbb{R}$ be defined by,

$$\phi(\mathcal{S};\mathcal{M},\mathbf{v}) = \frac{1}{2^{|\mathcal{F}\setminus\mathcal{S}|}} \sum\nolimits_{\mathbf{x} \in \Upsilon(\mathcal{S};\mathbf{v})} \kappa(\mathbf{x}) \tag{9}$$

Thus, given a set $\mathcal{S}$ of features, $\phi(\mathcal{S};\mathcal{M},\mathbf{v})$ represents the average value of the classifier over the points of feature space represented by $\Upsilon(\mathcal{S};\mathbf{v})$. The formulation presented in earlier work [Arenas et al. 2021a,b] allows for different input distributions when computing the average values. For the purposes of this paper, it suffices to consider solely a uniform input distribution, and so the dependency on the input distribution is not accounted for.

Table 2 illustrates how the average value is computed for two concrete sets of features. For example, if $\mathcal{S} = \{1,4\}$, then features 1 and 4 are fixed to value 0 (as dictated by $\mathbf{v}$).

We then allow all possible assignments to features 2 and 3, obtaining $\Upsilon(\{1,4\}) = \{(0,0,0,0),(0,0,1,0),(0,1,0,0),(0,1,1,0)\}$. To compute $\phi(\mathcal{S})$, we sum up the values of the rows of the truth table indicated by $\Upsilon(\mathcal{S})$, and divide by the total number of points, which is 4 in this case.

To simplify the notation, the following definitions are used throughout,

$$\Delta(i,\mathcal{S};\mathcal{M},\mathbf{v}) = (\phi(\mathcal{S}\cup\{i\};\mathcal{M},\mathbf{v}) - \phi(\mathcal{S};\mathcal{M},\mathbf{v})) \tag{10}$$

$$\varsigma(\mathcal{S};\mathcal{M},\mathbf{v}) = {}^{|\mathcal{S}|!(|\mathcal{F}|-|\mathcal{S}|-1)!}/_{|\mathcal{F}|!} \tag{11}$$

Finally, let $\mathsf{Sv} : \mathcal{F} \to \mathbb{R}$, i.e. the Shapley value for feature i, be defined by,

$$\mathsf{Sv}(i;\mathcal{M},\mathbf{v}) = \sum\nolimits_{\mathcal{S}\subseteq(\mathcal{F}\setminus\{i\})} \varsigma(\mathcal{S};\mathcal{M},\mathbf{v}) \times \Delta(i,\mathcal{S};\mathcal{M},\mathbf{v}) \tag{12}$$

Given an instance $(\mathbf{v}, c)$, the Shapley value assigned to each feature measures the contribution of that feature with respect to the prediction. A positive/negative value indicates that the feature can contribute to changing the prediction, whereas a value of 0 indicates no contribution.

Besides the polynomial-time computation of Shapley values for deterministic decomposable boolean circuits [Arenas et al. 2021b], for functions represented by truth tables, there exist polynomial-time algorithms (on the size of the truth table) for computing the Shapley values [Huang and Marques-Silva 2023]. This is illustrated in Figure 3.

It should be noted that both formal explanations (i.e. AXp's and CXp's) and Shapley values look at a prediction given a point $\mathbf{v}$ in feature space, but consider the function's behavior across all points that are consistent with the values dictated by $\mathbf{v}$. Additional similarities exist. Both formal explanations and Shapley values look at sets of features to fix, and then analyze all the points consistent with the fixed features. However, while formal explanations look at some of those sets of features to fix, Shapley values will analyze all possible subsets.

The use of Shapley values in explainability have been justified by significant claims. We illustrate some of the claims stated in earlier work [Strumbelj and Kononenko 2010]:

- "According to the 2nd axiom, if two features values have an identical influence on the prediction they are assigned contributions of equal size. The 3rd axiom says that if a feature has no influence on the prediction it is assigned a contribution of 0."
  (Note: the axioms above refer to the axiomatic characterization of Shapley values in [Strumbelj and Kononenko 2010].)
- "When viewed together, these properties ensure that any effect the features might have on the classifiers output will be reflected in the generated contributions, which effectively deals with the issues of previous general explanation methods."

Given the above, one would expect a direct correlation between a feature's importance and the absolute value of its Shapley value. As the rest of the paper shows, this is not the case.

3 Feature (Ir)relevancy

Given (6) and (7), we can aggregate the features that occur in AXp's and CXp's:

$$\mathcal{F}_{\mathbb{A}(\mathcal{E})} = \bigcup\nolimits_{\mathcal{X}\in\mathbb{A}(\mathcal{E})} \mathcal{X} \tag{13}$$

$$\mathcal{F}_{\mathbb{C}(\mathcal{E})} = \bigcup\nolimits_{\mathcal{Y}\in\mathbb{C}(\mathcal{E})} \mathcal{Y} \tag{14}$$

Moreover, MHS duality between the sets of AXp's and CXp's allows proving that: $\mathcal{F}_{\mathbb{A}(\mathcal{E})} = \mathcal{F}_{\mathbb{C}(\mathcal{E})}$. Hence, we just refer to $\mathcal{F}_{\mathbb{A}(\mathcal{E})}$ as the set of features that are contained in some AXp (or CXp).

A feature $i \in \mathcal{F}$ is relevant if it is contained in some AXp, i.e. $i \in \mathcal{F}_{\mathbb{A}(\mathcal{E})} = \mathcal{F}_{\mathbb{C}(\mathcal{E})}$; otherwise it is irrelevant, i.e. $i \notin \mathcal{F}_{\mathbb{A}(\mathcal{E})}$[3]. We will use the predicate $\mathsf{Relevant}(i)$ to denote that feature i is relevant, and predicate $\mathsf{Irrelevant}(i)$ to denote that feature i is irrelevant.

Relevant and irrelevant features provide a fine-grained characterization of feature importance, in that irrelevant features play no role whatsoever in prediction sufficiency. In fact, if $p \in \mathcal{F}$ is an irrelevant feature, then we can write:

$$\forall(\mathcal{X}\in\mathbb{A}(\mathcal{E})).\forall(u_p\in\mathcal{D}_p).\forall(\mathbf{x}\in\mathbb{F}).\\ \left[\bigwedge\nolimits_{i\in\mathcal{X}}(x_i = v_i)\land(x_p = u_p)\right] \rightarrow (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \tag{15}$$

The logic statement above clearly states that, if we fix the values of the features identified by any AXp then, no matter the value picked for feature p, the prediction is guaranteed to be $c = \kappa(\mathbf{v})$. The bottom line is that an irrelevant feature p is absolutely unimportant for the prediction, and so there is no reason to include it in a logic rule consistent with the instance.

For the example DT, we have that $\mathbb{A}(\mathcal{E}) = \{\{2,3,4\}\}$ and that $\mathbb{C}(\mathcal{E}) = \{\{2\},\{3\},\{4\}\}$, i.e. the explanation problem has one AXp and three CXp's. (Recall that the computation of both AXp's and CXp's is summarized in Figure 2.) As expected, $\mathcal{F}_{\mathbb{A}(\mathcal{E})} = \mathcal{F}_{\mathbb{C}(\mathcal{E})} = \{2,3,4\}$. Hence, we conclude that feature 1 is irrelevant, and that features 2, 3 and 4 are relevant. Observe that no AXp/CXp includes feature 1. For any AXp $\mathcal{X}$ this means that, adding feature 1 to $\mathcal{X}$, when feature 1 is assigned any value from its domain $\mathcal{D}_1$, would not change the prediction.

There are a few notable reasons for why irrelevant features are not considered in explanations. First, one can invoke Occam's razor (a mainstay of ML [Blumer et al. 1987]) and argue for simplest (i.e. irreducible) explanations. Second, if irreducibility of explanations were not a requirement, then one could claim that a prediction using all features would suffice, and that is never the case. Third, the fact that irrelevant features can take any value in their domain without that impacting the prediction shows how unimportant those features are.

[3] It should be noted that feature relevancy is tightly related with the concept of relevancy studied in logic-based abduction [Eiter and Gottlob 1995].

$\mathcal{S}$	rows for $\mathcal{S}$	rows for $\mathcal{S} \cup \{1\}$	$\phi(\mathcal{S})$	$\phi(\mathcal{S} \cup \{1\})$	$\Delta(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S}) \times \Delta(\mathcal{S})$
$\emptyset$	1..16	1..8	$3/16$	$3/8$	$3/16$	$0!(4-1)!/4! = 1/4$	$3/64$
$\{2\}$	1,2,3,4,9,10,11,12	1,2,3,4	$2/8$	$2/4$	$1/4$	$1!(4-2)!/4! = 1/12$	$1/48$
$\{3\}$	1,2,5,6,9,10,13,14	1,2,5,6	$2/8$	$2/4$	$1/4$	$1!(4-2)!/4! = 1/12$	$1/48$
$\{4\}$	1,3,5,7,9,11,13,15	1,3,5,7	$2/8$	$2/4$	$1/4$	$1!(4-2)!/4! = 1/12$	$1/48$
$\{2,3\}$	1,2,9,10	1,2	$1/4$	$1/2$	$1/4$	$2!(4-3)!/4! = 1/12$	$1/48$
$\{2,4\}$	1,3,9,11	1,3	$1/4$	$1/2$	$1/4$	$2!(4-3)!/4! = 1/12$	$1/48$
$\{3,4\}$	1,5,9,13	1,5	$1/4$	$1/2$	$1/4$	$2!(4-3)!/4! = 1/12$	$1/48$
$\{2,3,4\}$	1,9	1	0	0	0	$3!(4-4)!/4! = 1/4$	0
		Shapley value for feature 1				$\mathsf{Sv}(1)$ =	0.1719

$\mathcal{S}$	rows for $\mathcal{S}$	rows for $\mathcal{S} \cup \{2\}$	$\phi(\mathcal{S})$	$\phi(\mathcal{S} \cup \{2\})$	$\Delta(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S}) \times \Delta(\mathcal{S})$
$\emptyset$	1..16	1,2,3,4,9,10,11,12	$3/16$	$2/8$	$1/16$	$0!(4-1)!/4! = 1/4$	$1/64$
$\{1\}$	1..8	1,2,3,4	$3/8$	$2/4$	$1/8$	$1!(4-2)!/4! = 1/12$	$1/96$
$\{3\}$	1,2,5,6,9,10,13,14	1,2,9,10	$2/8$	$1/4$	0	$1!(4-2)!/4! = 1/12$	0
$\{4\}$	1,3,5,7,9,11,13,15	1,3,9,11	$2/8$	$1/4$	0	$1!(4-2)!/4! = 1/12$	0
$\{1,3\}$	1,2,5,6	1,2	$2/4$	$1/2$	0	$2!(4-3)!/4! = 1/12$	0
$\{1,4\}$	1,3,5,7	1,3	$2/4$	$1/2$	0	$2!(4-3)!/4! = 1/12$	0
$\{3,4\}$	1,5,9,13	1,9	$1/4$	0	$-1/4$	$2!(4-3)!/4! = 1/12$	$-1/48$
$\{1,3,4\}$	1,5	1	$1/2$	0	$-1/2$	$3!(4-4)!/4! = 1/4$	$-1/8$
		Shapley value for feature 2				$\mathsf{Sv}(2)$ =	-0.1198

$\mathcal{S}$	rows for $\mathcal{S}$	rows for $\mathcal{S} \cup \{3\}$	$\phi(\mathcal{S})$	$\phi(\mathcal{S} \cup \{3\})$	$\Delta(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S}) \times \Delta(\mathcal{S})$
$\emptyset$	1..16	1,2,5,6,9,10,13,14	$3/16$	$2/8$	$1/16$	$0!(4-1)!/4! = 1/4$	$1/64$
$\{1\}$	1..8	1,2,5,6	$3/8$	$2/4$	$1/8$	$1!(4-2)!/4! = 1/12$	$1/96$
$\{2\}$	1,2,3,4,9,10,11,12	1,2,9,10	$2/8$	$1/4$	0	$1!(4-2)!/4! = 1/12$	0
$\{4\}$	1,3,5,7,9,11,13,15	1,5,9,13	$2/8$	$1/4$	0	$1!(4-2)!/4! = 1/12$	0
$\{1,2\}$	1,2,3,4	1,2	$2/4$	$1/2$	0	$2!(4-3)!/4! = 1/12$	0
$\{1,4\}$	1,3,5,7	1,5	$2/4$	$1/2$	0	$2!(4-3)!/4! = 1/12$	0
$\{2,4\}$	1,3,9,11	1,9	$1/4$	0	$-1/4$	$2!(4-3)!/4! = 1/12$	$-1/48$
$\{1,2,4\}$	1,3	1	$1/2$	0	$-1/2$	$3!(4-4)!/4! = 1/4$	$-1/8$
		Shapley value for feature 3				$\mathsf{Sv}(3)$ =	-0.1198

$\mathcal{S}$	rows for $\mathcal{S}$	rows for $\mathcal{S} \cup \{4\}$	$\phi(\mathcal{S})$	$\phi(\mathcal{S} \cup \{4\})$	$\Delta(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S}) \times \Delta(\mathcal{S})$
$\emptyset$	1..16	1,3,5,7,9,11,13,15	$3/16$	$2/8$	$1/16$	$0!(4-1)!/4! = 1/4$	$1/64$
$\{1\}$	1..8	1,3,5,7	$3/8$	$2/4$	$1/8$	$1!(4-2)!/4! = 1/12$	$1/96$
$\{2\}$	1,2,3,4,9,10,11,12	1,3,9,11	$2/8$	$1/4$	0	$1!(4-2)!/4! = 1/12$	0
$\{3\}$	1,2,5,6,9,10,13,14	1,5,9,13	$2/8$	$1/4$	0	$1!(4-2)!/4! = 1/12$	0
$\{1,2\}$	1,2,3,4	1,3	$2/4$	$1/2$	0	$2!(4-3)!/4! = 1/12$	0
$\{1,3\}$	1,2,5,6	1,5	$2/4$	$1/2$	0	$2!(4-3)!/4! = 1/12$	0
$\{2,3\}$	1,2,9,10	1,9	$1/4$	0	$-1/4$	$2!(4-3)!/4! = 1/12$	$-1/48$
$\{1,2,3\}$	1,2	1	$1/2$	0	$-1/2$	$3!(4-4)!/4! = 1/4$	$-1/8$
		Shapley value for feature 4				$\mathsf{Sv}(4)$ =	-0.1198

Figure 3: Computation of Shapley values for the example DT and instance $((0,0,0,0),0)$. For each feature i, the sets to consider are all the sets that do not include the feature. For each set $\mathcal{S}$, we show the rows consistent with the values of the features in $\mathcal{S}$, as dictated by $\mathbf{v}$. For example, if $\mathcal{S} = \{2,4\}$, then the rows of the truth table consistent with having features 2 and 4 assigned value 0 are 1, 3, 9 and 11. The average values are obtained by summing up the values of the classifier in the rows consistent with $\mathcal{S}$ and dividing by the total number of rows. For $\mathcal{S} = \{2,4\}$, only row 3 in the truth table takes value 1, and so the average becomes $1/4$.

4 Refuting Shapley Values for Explainability

We now proceed to demonstrate that Shapley values for explainability can produce misleading information about feature importance, in that the relative feature importance obtained with Shapley values disagrees with the characterization of features in terms of (ir)relevancy. Clearly, information about feature (ir)relevancy is obtained by a rigorous, logic-based, analysis of the classifier, and so it captures precisely essential information about how the classifier's prediction depends (or not) on each of the features.

4.1 Misleading Feature Importance

Given the definition of Shapley values for explainability and of irrelevant features, we show that Shapley values will provide misleading information regarding relative feature importance, concretely that an irrelevant feature can be assigned the largest absolute Shapley value. Evidently, misleading information will cause human decision makers to consider features that are absolutely irrelevant for a prediction.

For the example DT, we have argued that feature 1 is irrelevant and that features 2, 3 and 4 are relevant. (The computation of AXp's and CXp's using a truth table is illustrated in Figure 2. Section 3 details how feature (ir)relevancy is decided.) Furthermore, from Figure 3, we obtain that,

$$\forall(i \in \{2,3,4\}).|\mathsf{Sv}(1)| > |\mathsf{Sv}(i)|$$

Thus, the feature with the largest absolute Shapley value is irrelevant for the prediction.

One might be tempted to argue that the sign of $\mathsf{Sv}(1)$ differs from the sign of $\mathsf{Sv}(2)$, $\mathsf{Sv}(3)$, $\mathsf{Sv}(4)$, and that that could explain the reported issue. However, the hypothetical relationship between the sign of the Shapley values and their perceived impact of the value of the prediction is a flawed argument, in that feature 1 plays no role in setting the prediction to 0, but feature 1 also plays no role in changing the value of the prediction. The results in the next section further confirm that the sign of a feature's Shapley value bears no direct influence on a (ir)relevancy of a feature.

4.2 Issues with Shapley Values for Explainability

By automating the analysis of boolean functions [Huang and Marques-Silva 2023], we have been able to identify a number of issues with Shapley values for explainability, all of which demonstrate that Shapley values can provide misleading information about the relative important of features. The list of possible issues is summarized in Table 3. Observe that some issues imply the occurrence of other issues, e.g. I4 implies I3, and I5 implies I2, among others. Our goal is to highlight a comprehensive range of problems that the use of Shapley values for explainability can induce, and so different issues aim to highlight such problems.

By analyzing all possible boolean functions defined on four variables, Table 4 summarizes the percentage of functions exhibiting the identified issues. For each possible function, the truth table for the function serves as the basis for the computation of all explanations, for deciding feature (ir)relevancy, and for the computation of Shapley values. The algorithms used are the ones sketched earlier in the paper, and all run in polynomial-time on the size of the truth table. For example, whereas issue I5, which is exemplified by the example DT and instance (also, see Section 4.1), occurs in 1.9% of the functions, issues I1, I2 and I6 occur in more than 55% of the functions, with I1 occurring in more than 99% of the functions. It should be noted that the identified issues were distributed evenly for instances where the prediction takes value 0 and instances where the prediction takes value 1. Moreover, it should be restated that the two constant functions were ignored.

4.3 Verdict & Justification

First, it should be plain that any of the issues described in Table 3 should be perceived as problematic in terms of assigning relative importance to features, with some issues serving to confirm the existence of misleading relative feature importance. This is the case with issues I2, I4, I5 and I7. However, assigning a Shapley value of 0 to a relevant feature or assigning non-zero Shapley value to an irrelevant feature will also cause human decision makers to overlook important features, or to analyze unimportant features. Such cases are also covered by the remaining issues.

Second, and as the results of the previous two sections amply demonstrate, the concept of Shapley values for explainability is fundamentally flawed. Furthermore, any explainability tool whose theoretical underpinnings are Shapley values for explainability is also fundamentally flawed.

Third, given the similarities between the computation of abductive and contrastive explanations and Shapley values for explainability, an immediate question is: why do Shapley values for explainability produce misleading measures of relative feature importance? It seems apparent that, whereas in the original definition of Shapley values for game theory, all coalitions are acceptable, this is not the case with explainability, i.e. some sets of features should not be considered when assigning importance to a feature. Thus, one reason that causes Shapley values to produce misleading information is the fact that some disallowed set of features are accounted for.

5 Discussion

This paper presents a simple argument demonstrating that Shapley values for explainability can produce misleading information regarding relative feature importance. A number of potential issues with Shapley values for explainability has been identified, and shown to occur rather frequently in boolean classifiers. It is therefore plain that the continued use of XAI approaches based on Shapley values (see [Huang and Marques-Silva 2023] for additional references) in high-risk domains will inevitably cause human decision makers to assign importance to unimportant features, and to overlook important features. Evidently, such uses of Shapley values for explainability are bound to have unwanted grave consequences.

Issue	Condition
I1	$\exists(i \in \mathcal{F}).[\mathsf{Irrelevant}(i) \land (\mathsf{Sv}(i) \neq 0)]$
I2	$\exists(i_1, i_2 \in \mathcal{F}).[\mathsf{Irrelevant}(i_1) \land \mathsf{Relevant}(i_2) \land (\lvert\mathsf{Sv}(i_1)\rvert > \lvert\mathsf{Sv}(i_2)\rvert)]$
I3	$\exists(i \in \mathcal{F}).[\mathsf{Relevant}(i) \land (\mathsf{Sv}(i) = 0)]$
I4	$\exists(i_1, i_2 \in \mathcal{F}).[\mathsf{Irrelevant}(i_1) \land (\mathsf{Sv}(i_1) \neq 0)] \land [\mathsf{Relevant}(i_2) \land (\mathsf{Sv}(i_2) = 0)]$
I5	$\exists(i \in \mathcal{F}).[\mathsf{Irrelevant}(i) \land \forall(1 \leq j \leq m, j \neq i).\lvert\mathsf{Sv}(i)\rvert > \lvert\mathsf{Sv}(j)\rvert]$
I6	$\exists(i_1, i_2 \in \mathcal{F}).[\mathsf{Irrelevant}(i_1) \land \mathsf{Relevant}(i_2) \land (\mathsf{Sv}(i_1) \times \mathsf{Sv}(i_2) > 0)]$
I7	$\exists(i_1, i_2 \in \mathcal{F}).[\mathsf{Irrelevant}(i_1) \land \mathsf{Relevant}(i_2) \land (\lvert\mathsf{Sv}(i_1)\rvert > \lvert\mathsf{Sv}(i_2)\rvert) \land (\mathsf{Sv}(i_1) \times \mathsf{Sv}(i_2) > 0)]$

Table 3: Identified potential issues with Shapley values

Metric	Value
# of functions	65534
# number of instances	1048544
# of I1 issues	781696
# of functions exhibiting I1 issues	65320
% functions exhibiting I1 issues	99.67
# of I2 issues	105184
# of functions exhibiting I2 issues	40448
% functions exhibiting I2 issues	61.72
# of I3 issues	43008
# of functions exhibiting I3 issues	7800
% functions exhibiting I3 issues	11.90
# of I4 issues	5728
# of functions exhibiting I4 issues	2592
% functions exhibiting I4 issues	3.96
# of I5 issues	1664
# of functions exhibiting I5 issues	1248
% functions exhibiting I5 issues	1.90
# of I6 issues	109632
# of functions exhibiting I6 issues	36064
% functions exhibiting I6 issues	55.03
# of I7 issues	11776
# of functions exhibiting I7 issues	7632
% functions exhibiting I7 issues	11.65

Table 4: Results over all 4-variable boolean functions. The two constant functions were discarded, since κ is required not to be constant.

If exact computation of Shapley values for explainability yields misleading information regarding relative feature importance then, evidently, any XAI tool that claims to approximate Shapley values for explainability, e.g. SHAP and variants [Chen et al. 2022; Lundberg et al. 2020; Lundberg and Lee 2017], cannot guarantee not to produce misleading relative feature importance.

More importantly, it should be underlined that our recent experimental results [Huang and Marques-Silva 2023] suggest little to no correlation between exact Shapley values and the results produced by SHAP [Lundberg and Lee 2017]. To put it bluntly, a flawed approximation of a flawed concept does not offer any guarantees whatsoever regarding the rigor of that flawed approximation at estimating feature attribution values. Evidently, the continued practical use of tools that approximate Shapley values is also bound to have unwanted grave consequences.

Given the demonstrated inadequacy of Shapley values for explainability, a natural line of research is whether an alternative metric could be devised which respects feature (ir)relevancy. Although recent work proposed a possible metric [Huang and Marques-Silva 2023], it is unclear how it could be related with Shapley values for explainability.

Acknowledgments

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". This work was motivated in part by discussions with several colleagues including L. Bertossi, A. Ignatiev, N. Narodytska, M. Cooper, Y. Izza, R. Passos, J. Planes and N. Asher.

References

Marcelo Arenas, Pablo Barceló, Leopoldo E. Bertossi, and Mikaël Monet. 2021a. On the Complexity of SHAP-Score-Based Explanations: Tractability via Knowledge Compilation and Non-Approximability Results. CoRR abs/2104.08015 (2021). arXiv:2104.08015 https://arxiv.org/abs/2104.08015

Marcelo Arenas, Pablo Barceló, Leopoldo E. Bertossi, and Mikaël Monet. 2021b. The Tractability of SHAP-Score-Based Explanations for Classification over Deterministic and Decomposable Boolean Circuits. In AAAI. 6670–6678.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one 10, 7 (2015), e0130140.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1987. Occam's Razor. Inf. Process. Lett. 24, 6 (1987), 377–380. https://doi.org/10.1016/0020-0190(87)90114-1

Hugh Chen, Ian C. Covert, Scott M. Lundberg, and Su-In Lee. 2022. Algorithms to estimate Shapley value feature attributions. CoRR abs/2207.07605 (2022). https://doi.org/10.48550/arXiv.2207.07605 arXiv:2207.07605

Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2019. L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. In ICLR.

Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In IEEE S&P. 598–617.

Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu. 2021. On the Tractability of SHAP Explanations. In AAAI. 6505–6513.

Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu. 2022. On the Tractability of SHAP Explanations. J. Artif. Intell. Res. 74 (2022), 851–886. https://doi.org/10.1613/jair.1.13283

Thomas Eiter and Georg Gottlob. 1995. The Complexity of Logic-Based Abduction. J. ACM 42, 1 (1995), 3–42. https://doi.org/10.1145/200836.200838

Xuanxiang Huang and Joao Marques-Silva. 2023. The Inadequacy of Shapley Values for Explainability. CoRR abs/2302.08160 (2023). https://doi.org/10.48550/arXiv.2302.08160 arXiv:2302.08160

Alexey Ignatiev. 2020. Towards Trustable Explainable AI. In IJCAI. 5154–5158.

Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva. 2020. From Contrastive to Abductive Explanations and Back Again. In AIxIA. 335–355.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. 2019a. Abduction-Based Explanations for Machine Learning Models. In AAAI. 1511–1519.

Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. 2019b. On Validating, Repairing and Refining Heuristic ML Explanations. CoRR abs/1907.02509 (2019). arXiv:1907.02509 http://arxiv.org/abs/1907.02509

Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. 2022. On Tackling Explanation Redundancy in Decision Trees. J. Artif. Intell. Res. 75 (2022), 261–321. https://jair.org/index.php/jair/article/view/13575/

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle A. Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. In ICML. 5491–5500.

Stan Lipovetsky and Michael Conklin. 2001. Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry 17, 4 (2001), 319–330.

Scott M. Lundberg, Gabriel G. Erion, Hugh Chen, Alex J. DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 1 (2020), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In NeurIPS. 4765–4774.

Joao Marques-Silva. 2022. Logic-Based Explainability in Machine Learning. In Reasoning Web. 24–104.

Joao Marques-Silva. 2023. Disproving XAI Myths with Formal Methods – Initial Results. In ICECCS.

Joao Marques-Silva and Alexey Ignatiev. 2022. Delivering Trustworthy AI through Formal XAI. In AAAI. 12342–12350.

Luke Merrick and Ankur Taly. 2020. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. In CDMAKE. 17–38.

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

Nina Narodytska, Aditya A. Shrotri, Kuldeep S. Meel, Alexey Ignatiev, and Joao Marques-Silva. 2019. Assessing Heuristic Machine Learning Explanations with Model Counting. In SAT. 267–278.

Raymond Reiter. 1987. A Theory of Diagnosis from First Principles. Artif. Intell. 32, 1 (1987), 57–95. https://doi.org/10.1016/0004-3702(87)90062-2

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In KDD. 1135–1144.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In AAAI. 1527–1535.

Alvin E Roth. 1988. The Shapley value: essays in honor of Lloyd S. Shapley. Cambridge University Press.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (Eds.). 2019. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer. https://doi.org/10.1007/978-3-030-28954-6

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

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

Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2021. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. In NeurIPS. 9391–9404.

Erik Strumbelj and Igor Kononenko. 2010. An Efficient Explanation of Individual Classifications using Game Theory. J. Mach. Learn. Res. 11 (2010), 1–18. https://dl.acm.org/doi/10.5555/1756006.1756007

Erik Strumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41, 3 (2014), 647–665. https://doi.org/10.1007/s10115-013-0679-x

David S. Watson. 2022. Rational Shapley Values. In FAccT. 1083–1094.