DOI:10.1145/3635301

When the decisions of ML models impact people, one should expect explanations to offer the strongest guarantees of rigor. However, the most popular XAI approaches offer none.

BY JOAO MARQUES-SILVA AND XUANXIANG HUANG

Explainability Is *Not* a Game

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[3,23,30,31,33] offer no guarantees of rigor. Unsurprisingly, a number of works have demonstrated several misconceptions of informal approaches to XAI.[13,16,25] In contrast to informal XAI, formal explainability offers a logic-based, model-precise approach for computing explanations.[18] 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.[24,26]

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[34] were originally proposed in the context of game theory, but have found a wealth of application domains.[32] More importantly, for more than two decades, Shapley values have been proposed in the context of explaining the decisions of complex ML models,[7,21,23,37,38] with their use in XAI being referred to as *SHAP scores*. The importance of SHAP scores for explainability is illustrated by the massive impact of tools like SHAP,[23] including many recent uses that have a direct influence on human beings (see Huang and Marques-Silva[13] for some recent references).

Due to their computational complexity, the exact computation of SHAP scores has not been studied in practice. Hence, it is unclear how good existing approximate solutions are, with a well-known example being SHAP.[5,22,23] Recent work[1] proposed a polynomial-time algorithm for computing SHAP scores 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[23] with those obtained with exact algorithms.

» key insights

- **Shapley values find extensive uses in explaining machine learning models and serve to assign importance to the features of the model.**
- **Shapley values for explainability also find ever-increasing uses in high-risk and safety-critical domains, for example, medical diagnosis.**
- **This article proves that the existing definition of Shapley values for explainability can produce misleading information regarding feature importance, and so can induce human decision makers in error.**

IMAGE BY PHOENIX 1319

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

A number of earlier works reported practical problems with explainability approaches that use SHAP scores[20] (Huang and Marques-Silva[13] covers a number of additional references). However, these works focus on practical tools, which approximate SHAP scores, but do not investigate the possible existence of fundamental limitations with the existing definitions of SHAP scores. In contrast with these other works, this article exhibits simple classifiers for which relative feature importance obtained with SHAP scores provides misleading information, in that features that bear no significance for a prediction can be deemed more important, in terms of SHAP scores, than features that bear some significance for the same prediction. The importance of this article's results, and of the identified limitations of SHAP scores, should be assessed in light of the fast-growing uses of explainability solutions in domains that directly impact human beings, for example, medical diagnostic applications, especially when the vast majority of such uses build on SHAP scores.[a]

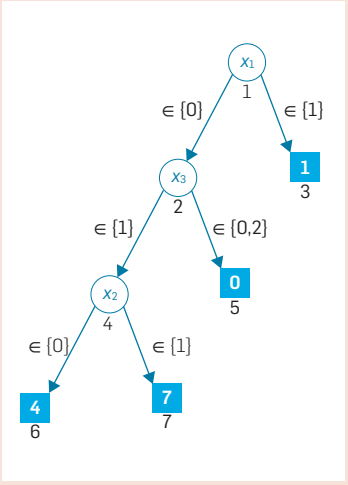
a Given the many envisioned applications of ML, our claim is that explainability is a topic of such importance that it should not be gambled upon. The main conclusion to draw from our work is that the existing foundations of SHAP scores must be revised.

The article starts by introducing the notation and definitions used throughout as well as two example decision trees. Afterward, the article presents a brief introduction to formal explanations and adversarial examples[11] along with the concepts of relevancy/irrelevancy, which have been studied in logic-based abduction since the mid 1990s.[10] The article then overviews SHAP scores and illustrates their computation for the running examples. This will serve to show that SHAP scores can provide misleading information regarding relative feature importance. To conclude, we extend further the negative results regarding the inadequacy of using SHAP scores for feature attribution in XAI and briefly summarize our main results.

Definitions

Throughout the article, we adopt the notation and the definitions in-

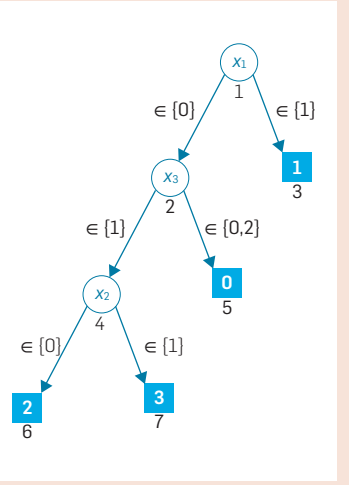
Figure 1. Decision trees and their tabular representations. For these two classifiers, we have $\mathcal{F} = \{1,2,3\}$, $\mathbb{D}_1 = \mathbb{D}_2 = \{0,1\}$, and $\mathbb{D}_3 = \{0,1,2\}$, $\mathbb{F} = \{0,1\}^2 \times \{0,1,2\}$, and $\mathcal{K} = \{0,1,2,3,4,5,6,7\}$, albeit the DTs and TRs only use a subset of the classes. Literals in the DTs are represented with set notation, to enable more compact DTs. The classification functions are given by the decision trees, or alternatively by the tabular representations. The instance considered is $((1,1,2),1)$, which is consistent with the highlighted path $\langle 1,3\rangle$ in both DTs. The prediction is 1, as indicated in terminal node 3.



(a) Decision tree (DT) for κ_1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

(b) Tabular representations (TRs) for κ_1 and κ_2



(c) Decision tree (DT) for κ_2

Analysis of the two classifiers:

By inspection of the DTs/TRs for both κ_1 and κ_2, it is immediate that: (i) if $x_1 = 1$, then the prediction must be 1; (ii) if $x_1 = 0$, then the prediction cannot be 1. Thus, the classifiers predict class 1 (or predict a class other than 1) *independently* of the values assigned to x_2 and x_3. Hence, for point $\mathbf{v} = (1, 1, 2)$, the prediction will be 1 as long as $x_1 = 1$, *independently* of x_2 and x_3. To change the prediction from class 1, the value of x_1 must be changed. In this case, the prediction will change to a class other than 1 independently of the values assigned to x_2 and x_3. These observations apply to both κ_1 and κ_2. The influence of features can be related with adversarial examples for the prediction of class 1. To change the prediction, it must be the case that feature 1 changes its value. Also, any l_0 minimal adversarial example will only change the value of feature 1.

(d) Influence of features on κ_1 and κ_2 for ((1, 1, 2), 1)

troduced in earlier work, namely Marques-Silva[24,26] and Arenas et al.[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 $\mathbb{D}_i$. Domains can be ordinal (for example, real- or integer-valued) or categorical. Feature space is defined by the cartesian product of the domains of the features: $\mathbb{F} = \mathbb{D}_1 \times \cdots \times \mathbb{D}_m$. A classifier $\mathcal{M}$ computes a (non-constant) classification function:[b] $\kappa: \mathbb{F} \to \mathcal{K}$. A classifier $\mathcal{M}$ is associated with a tuple $(\mathcal{F}, \mathbb{F}, \mathcal{K}, \kappa)$. For the purposes of this article, we restrict the features' domains to be discrete. 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 $(\mathcal{M}, (\mathbf{v}, c))$. Thus, $\mathbf{v}$ represents a concrete point in feature space, whereas $\mathbf{x} \in \mathbb{F}$ represents an arbitrary point in feature space.

Running examples. The two decision trees (DTs) shown in Figures 1a and 1c represent the running examples used throughout the article. A tabular representation of the two DTs is depicted in Figure 1b. A classification function κ_1 is associated with the first DT and a classification function κ_2 is associated with the second DT. Given the information shown in the DT, we have that $\mathcal{F} = \{1,2,3\}$, $\mathbb{D}_i = \{0,1\}, i = 1,2$, $\mathbb{D}_3 = \{0,1,2\}$, $\mathbb{F} = \{0,1\}^2 \times \{0,1,2\}$, and $\mathcal{K} = \{0,1,2,3,4,5,6,7\}$, but some classes are not used by some of the DTs. Throughout the article, we consider exclusively the instance $(\mathbf{v}, c) = ((1,1,2),1)$. Figure 1d discusses the role of each feature in predicting class 1, and in predicting some other class. As argued, for predicting class 1, only the value of feature 1 matters. For predicting a class other than 1, only the value of feature 1 matters. Clearly, the values of the other features matter to decide which actual class other than 1

b A classifier that computes a constant function, that is, the same prediction for all points in feature space, is uninteresting, and so it is explicitly disallowed.

is picked. However, given the target instance, our focus is to explain the predicted class 1.

The remainder of the article answers the following questions regarding the instance $((1,1,2),1)$. How does the intuitive analysis of Figure 1d relate with formal explanations? How does it relate with adversarial examples? How does it relate with SHAP scores? What can be said about the relative feature importance determined by SHAP scores? For the two fairly similar DTs, and given the specific instance considered, are there observable differences in the computed explanations, adversarial examples, and SHAP scores? Are there observable differences, between the two DTs, regarding relative feature importance as determined by SHAP scores?

A hypothetical scenario. To motivate the analysis of the classifiers in Figure 1, we consider the following hypothetical scenario.[c] A small college aims to predict the number of extra-curricular activities of each of its students, where this number can be between 0 and 7. Let feature 1 represent whether the student is an honors student (0 for no, and 1 for yes). Let feature 2 represent where the student originates from, that is, an urban or non-urban household (0 for non-urban, and 1 for urban). Finally, let feature 3 represent whether the student's field of study is humanities, arts, or sciences (0 for humanities, 1 for arts, 2 for sciences). Thus, the target instance $((1,1,2),1)$ denotes an honors student from an urban household, studying sciences, for whom the predicted number of extra-curricular activities is 1.

Parameterized example. In the following sections, we consider a more general parameterized classifier, shown in Table 1, which encompasses the two classifiers shown in Figure 1. As clarified below, we impose that,

$$\alpha \neq \sigma_j, j = 1\ldots,6 \tag{1}$$

For simplicity, we also require that $\alpha, \sigma_j \in \mathbb{Z}, j = 1,\ldots,6$. It is plain that the DT of Figure 1a is a concrete instantiation of the parameterized classifier shown in Table 1, by setting $\sigma_1 = \sigma_3 = \sigma_4 = \sigma_6 = 0$, $\sigma_2 = 4$, $\sigma_5 = 7$ and $\alpha = 1$. For the parameterized example, the instance to be considered is $((1,1,2),\alpha)$.

c It would be fairly straightforward to create two datasets, for example, the two TRs shown, from which the DTs shown in Figures 1a and 1c would be induced using existing tree learning tools.

Features that bear no significance for a prediction can be deemed more important, in terms of SHAP scores, than features that bear some significance for the same prediction.

Formal Explanations and Adversarial Examples

Formal explanations. We follow recent accounts of formal explainability.[24] In the context of XAI, abductive explanations (AXps) have been studied since 2018.[18,35] Similar to other heuristic approaches, for example, Anchors,[31] abductive explanations are an example of explainability by feature selection, that is, a subset of features is selected as the explanation. AXps represent a rigorous example of explainability by feature selection, and can be viewed as the answer to a "Why (the prediction, given v)?" 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:

$$\forall(\mathbf{x} \in \mathbb{F}).\left[\bigwedge\nolimits_{i\in\mathcal{X}}(x_i = v_i)\right] \rightarrow (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \tag{2}$$

For simplicity, we associate a predicate WAXp (that is, weak AXp) with (2), such that $\mathtt{WAXp}(\mathcal{X};\mathscr{E})$ holds if and only if (2) holds.

Observe that (2) is monotone on $\mathcal{X}$,[24]

Table 1. Tabular representation of parameterized classifier, κ.

row #	x_1	x_2	x_3	κ(x)
1	0	0	0	σ_1
2	0	0	1	σ_2
3	0	0	2	σ_3
4	0	1	0	σ_4
5	0	1	1	σ_5
6	0	1	2	σ_6
7	1	0	0	α
8	1	0	1	α
9	1	0	2	α
10	1	1	0	α
11	1	1	1	α
12	1	1	2	α

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

$$\texttt{WAXp}(\mathcal{X};\mathcal{E}) \land \forall(t \in \mathcal{X}).\neg\texttt{WAXp}(\mathcal{X}\backslash\{t\};\mathcal{E}) \quad (3)$$

A predicate $\texttt{AXp}:2^{\mathcal{F}} \rightarrow \{0,1\}$ is associated with (3), such that $\texttt{AXp}(\mathcal{X};\mathcal{E})$ holds true if and only if (3) holds true.[d]

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

$$\text{IF } \left[\bigwedge\nolimits_{i\in\mathcal{X}}(x_i = v_i)\right] \text{ THEN } (\kappa(\mathbf{x}) = c) \quad (4)$$

where $c = \kappa(\mathbf{v})$. It should be noted that informal XAI methods have also proposed the use of IF-THEN rules[31] which, in the case of Anchors[31] may or may not be sound.[16,18] In contrast, rules obtained from AXps are logically sound.

Moreover, contrastive explanations (CXps) represent a type of explanation that differs from AXps, in that CXps answer a "Why Not (some other prediction, given v)?" question.[17,28] Given a set $\mathcal{Y} \subseteq \mathcal{F}$, sufficiency for changing the prediction can be stated formally:

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

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 prediction can be changed by choosing suitable values to those features. As shown, and for simplicity, we associate a predicate $\texttt{WCXp}$ (that is, weak CXp) with (5), such that $\texttt{WCXp}(\mathcal{Y};\mathcal{E})$ holds if and only if (5) holds.

Similarly to the case of AXps, for CXps (5) is monotone on $\mathcal{Y}$, and so the two conditions (sufficiency for changing the prediction and subset-minimality) can be stated formally as follows:

$$\texttt{WCXp}(\mathcal{Y};\mathcal{E}) \land \forall(t \in \mathcal{Y}).\neg\texttt{WCXp}(\mathcal{Y}\backslash\{t\};\mathcal{E}) \quad (6)$$

A predicate $\texttt{CXp} : 2^{\mathcal{F}} \rightarrow \{0,1\}$ is associated with (6), such that $\texttt{CXp}(\mathcal{Y};\mathcal{E})$ holds true if and only if (6) holds true.

Algorithms for computing AXps and CXps for different families of classifiers have been proposed in recent years (Marques-Silva[26] provides a recent account of the progress observed in computing formal explanations). These algorithms include the use of automated reasoners (for example, 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 AXps and CXps are represented by:

$$\mathbb{A}(\mathcal{E}) = \{\mathcal{X} \subseteq \mathcal{F} \mid \texttt{AXp}(\mathcal{X};\mathcal{E})\} \quad (7)$$
$$\mathbb{C}(\mathcal{E}) = \{\mathcal{Y} \subseteq \mathcal{F} \mid \texttt{CXp}(\mathcal{Y};\mathcal{E})\} \quad (8)$$

For example, $\mathbb{A}(\mathcal{E})$ represents the set of all logic rules (of the type of rule (4)) that predict $c = \kappa(\mathbf{v})$, which are consistent with $\mathbf{v}$, and which are irreducible (that is, no literal $x_i = v_i$ can be discarded).

Furthermore, it has been proved[17] 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 CXps; and (ii) a set $\mathcal{Y} \subseteq \mathcal{F}$ is a CXp if and only if it is an MHS of the set of AXps. This property is referred to as MHS duality, and can be traced back to the seminal work of R. Reiter[29] in model-based diagnosis. Moreover, MHS duality has been shown to be instrumental for the enumeration of AXps and CXps, but also for answering other explainability queries.[24] Formal explainability has made significant progress in recent years, covering a wide range of topics of research.[24,26]

Feature (ir)relevancy. Given (7) and (8), we can aggregate the features that occur in AXps and CXps:

$$\mathcal{F}_{\mathbb{A}(\mathcal{E})} = \bigcup\nolimits_{\mathcal{X}\in\mathbb{A}(\mathcal{E})}\mathcal{X} \;;\; \mathcal{F}_{\mathbb{C}(\mathcal{E})} = \bigcup\nolimits_{\mathcal{Y}\in\mathbb{C}(\mathcal{E})}\mathcal{Y} \quad (9)$$

Moreover, MHS duality between the sets of AXps and CXps allows one to prove 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, that is, $i \in \mathcal{F}_{\mathbb{A}(\mathcal{E})} = \mathcal{F}_{\mathbb{C}(\mathcal{E})}$; otherwise it is irrelevant, that is, $i \notin \mathcal{F}_{\mathbb{A}(\mathcal{E})}$.[e] A feature that occurs in all AXps is referred to as *necessary*. We will use the predicate $\texttt{Relevant}(i)$ to denote that feature i is relevant, and predicate

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

e It should be noted that feature relevancy is tightly related with the concept of relevancy studied in logic-based abduction.[10]

Figure 2. Computing AXps/CXps for the example parameterized classifier shown in Table 1 and instance $(\mathbf{v},c) = ((1,1,2),\alpha)$. All subsets of features are considered. For computing AXps/CXps, and for some set $\mathcal{X}$, the features in $\mathcal{X}$ are fixed to their values as determined by v. The picked rows, that is, $\texttt{rows}(\mathcal{X})$, are the rows consistent with those fixed values.

$\mathcal{S}$	rows($\mathcal{S}$)	WAXp($\mathcal{S}$)? $\mathcal{S}$ sufficient?	AXp($\mathcal{S}$)? $\mathcal{S}$ also minimal?	$\mathcal{F}\backslash\mathcal{S}$	rows($\mathcal{F}\backslash\mathcal{S}$)	WCXp($\mathcal{S}$)? $\mathcal{S}$ changed κ?	CXp($\mathcal{S}$)? $\mathcal{S}$ also minimal?
∅	1...12	×		{1, 2, 3}	12	×	×
{1}	7,8,9,10,11,12	✓	✓	{2, 3}	6,12	✓	✓
{2}	4,5,6,10,11,12	×		{1, 3}	9,12	×	
{3}	3,6,9,12	×		{1, 2}	10,11,12	×	
{1, 2}	10,11,12	✓	×	{3}	3,6,9,12	✓	×
{1, 3}	9,12	✓	×	{2}	4,5,6,10,11,12	✓	×
{2, 3}	6,12	×		{1}	7,8,9,10,11,12	×	
{1, 2, 3}	12	✓	×	∅	1...12	✓	×

Irrelevant(i) to denote that feature i is irrelevant.

Relevant and irrelevant features provide a coarse-grained characterization of feature importance, in that irrelevant features play no role whatsoever in prediction sufficiency. In fact, the following logic statement is true iff $r \in \mathcal{F}$ is an irrelevant feature:

$$\forall(\mathcal{X} \in \mathbb{A}(\mathcal{E})).\forall(u_r \in \mathbb{D}_r).\forall(\mathbf{x} \in \mathbb{F}). \left[\bigwedge_{i \in \mathcal{X}\setminus\{r\}} (x_i = v_i) \land (x_r = u_r)\right] \rightarrow (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \quad (10)$$

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

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[4]) and argue for simplest (that is, irreducible) explanations. Second, if irreducibility of explanations were not a requirement, then one could claim that a prediction using all the features would suffice, but that is never of interest. Third, the fact that irrelevant features can take *any* value in their domain without impacting the value of the prediction shows how unimportant those features are.

Explanations for running examples. For both the classifiers of Figure 1 or the parameterized classifier of Table 1 (under the stated assumptions), it is simple to show that $\mathbb{A} = \{\{1\}\}$ and that $\mathbb{C} = \{\{1\}\}$. The computation of AXps/CXps is shown in Figure 2 for the parameterized classifier. More detail on how Figure 2 is obtained is given in Table 2. (Observe that computation of AXps/CXps shown holds as long as α *differs* from each of the σ_i, $i = 1,\ldots,6$, that is, (1) holds.) Thus, for any of these classifiers, feature 1 is relevant (in fact it is necessary), and features 2 and 3 are irrelevant. These results agree with the analysis of the classifier (see Figure 1d) in terms of feature influence, in that feature 1 occurs in explanations, and features 2 and 3 do not. More importantly, there is no difference in terms of the computed explanations for either κ_1 or κ_2.

Table 2. Examples of how each set is analyzed when computing AXps. It is assumed that $\alpha \neq \sigma_i$, $i = 1, \ldots, 6$. For CXps, a similar approach is used.

$\mathcal{S}$	Template	Rows	$\mathbf{x}$	$\kappa(\mathbf{x})$	$\forall(\mathbf{x} \in \mathbb{F}).(\kappa(\mathbf{x}) = c)$?
{3}	$(x_1, x_2, 2)$	3	(0, 0, 2)	σ_3	No
		6	(0, 1, 2)	σ_6	
		9	(1, 0, 2)	α	
		12	(1, 1, 2)	α	
{1, 2}	$(1, 1, x_3)$	10	(1, 1, 0)	α	Yes
		11	(1, 1, 1)	α	
		12	(1, 1, 2)	α	
{2, 3}	$(x_1, 1, 2)$	6	(0, 1, 2)	σ_6	No
		12	(1, 1, 2)	α	

As can be concluded, the computed abductive and contrastive explanations agree with the analysis shown in Figure 1d in terms of feature influence. Indeed, features 2 and 3, which have no influence in determining nor in changing the predicted class 1, are not included in the computed explanations, that is, they are irrelevant. In contrast, feature 1, which is solely responsible for the prediction, is included in the computed explanations, that is, it is relevant. Also unsurprisingly,[12,17,19] the existence of adversarial examples is tightly related with CXps, and so indirectly with AXps. These observations should be expected, since abductive explanations are based on logic-based abduction, and so have a precise logic formalization.

Adversarial examples. Besides studying the relationship between SHAP scores and formal explanations, we also study their relationship with adversarial examples in ML models.[11]

Hamming distance is a measure of distance between points in feature space. The Hamming distance is also referred to as the l_0 measure of distance, and it is defined as follows:

$$||\mathbf{x} - \mathbf{y}||_0 = \sum_{i=1}^{m} \text{ITE}(x_i \neq y_i, 1, 0) \quad (11)$$

Given a point $\mathbf{v}$ in feature space, an *adversarial example* (AE) is some other point $\mathbf{x}$ in feature space that changes the prediction and such that the measure of distance l_p between the two points is small enough:

$$||\mathbf{x} - \mathbf{v}||_p \leq \varepsilon \land (\kappa(\mathbf{x}) \neq \kappa(\mathbf{v})) \quad (12)$$

(in our case, we consider solely $p = 0$.)

The relationship between adversarial examples and explanations is well-known.[12,17,19]

Proposition 1.
If an explanation problem $\mathcal{E} = (\mathcal{M}, (\mathbf{v}, c))$ has a CXp $\mathcal{Y}$, then the classifier $\mathcal{M}$ on instance $(\mathbf{v}, c)$ has an adversarial example $\mathbf{z}$, with $||\mathbf{z} - \mathbf{v}||_0 = |\mathcal{Y}|$.

Similarly, it is straightforward to prove that,

Proposition 2.
If a classifier $\mathcal{M}$ on instance $(\mathbf{v}, c)$ has an adversarial example with l_0 distance δ that includes an irrelevant feature $j \in \mathcal{F}$, then there exists an adversarial example with l_0 distance $\delta - 1$ that does not include j.

Thus, irrelevant features are not included in subset- (or cardinality-) minimal adversarial examples.

Adversarial examples for running examples. Given the previous discussion, and for $(\mathbf{v}, c) = ((1, 1, 2), \alpha)$, it is plain that there exists an adversarial example with a l_0 distance of 1, which consists of changing the value of feature 1 from 1 to 0. This adversarial example is both subset-minimal and cardinality-minimal. As with formal explanations, the minimal adversarial example agrees with the analysis of the two classifiers included in Figure 1d.

SHAP Scores

Shapley values were proposed in the 1950s, in the context of game theory,[34]

Table 3. Computation of average values. Column 'Rows' shows the row numbers in Table 1 to consider when computing the average value.

$\mathcal{S}$	Template	$\Upsilon(\mathcal{S};\mathbf{v})$	Rows	$\phi(\mathcal{S})$
{1, 2}	(1, 1, x_3)	{(1, 1, 0), (1, 1, 1), (1, 1, 2)}	{10, 11, 12}	α
{2, 3}	(x_1, 1, 2)	{(0, 1, 2), (1, 1, 2)}	{6, 12}	$\sigma_6/2 + \alpha/2$

and find a wealth of uses.[32] More recently, Shapley values have been extensively used for explaining the predictions of ML models.[6,7,21,23,27,36,37,38,39] among a vast number of recent examples (see Huang and Marques-Silva[13] for a more comprehensive list of references), and are commonly referred to as SHAP scores.[23] SHAP scores represent one example of explainability by feature attribution, that is, some score is assigned to each feature as a form of explanation. Moreover, their complexity has been studied in recent years.[1,2,8,9] This section provides a brief overview of how SHAP scores are computed. Throughout, we build on the notation used in recent work,[1,2] which builds on the work of Lundberg and Lee.[23]

Computing SHAP scores. Let Υ: $2^{\mathcal{F}} \rightarrow 2^{\mathbb{F}}$ be defined by,

$$\Upsilon(\mathcal{S};\mathbf{v}) = \left\{\mathbf{x} \in \mathbb{F} \;\middle|\; \wedge_{i\in\mathcal{S}} x_i = v_i\right\} \quad (13)$$

that is, 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 we are interested in when computing explanations (see Figure 2).

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

$$\phi(\mathcal{S};\mathscr{M},\mathbf{v}) = \frac{1}{\Pi_{i\in\mathcal{F}\setminus\mathcal{S}}|\mathbb{D}_i|}\sum\nolimits_{\mathbf{x}\in\Upsilon(\mathcal{S};\mathbf{v})}\kappa(\mathbf{x}) \quad (14)$$

Thus, given a set $\mathcal{S}$ of features, $\phi(\mathcal{S};\mathscr{M},\mathbf{v})$ 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[1,2] allows for different input distributions when computing the average values. For the purposes of this article, it suffices to consider solely a uniform input distribution, and so the dependency on the input distribution is not accounted for.

Table 3 illustrates how the average value is computed for two concrete sets of features. As an example, if $\mathcal{S} = \{1\}$, then feature 1 is fixed to value 1 (as dictated by $\mathbf{v}$). We then allow all possible assignments to features 2 and 3, thus obtaining,

$$\Upsilon(\{1\}) = \{\ (1,0,0),(1,0,1),\ (1,0,2),(1,1,0),\ (1,1,1),(1,1,2)\}$$

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 6 in this case.

To simplify the notation, the following definitions are used,

$$\Delta(i,\mathcal{S};\mathscr{M},\mathbf{v}) = ((\phi(\mathcal{S}\cup\{i\};\mathscr{M},\mathbf{v}) - \phi(\mathcal{S};\mathscr{M},\mathbf{v})) \quad (15)$$

$$\varsigma(\mathcal{S};\mathscr{M},\mathbf{v}) = {}^{|\mathcal{S}|!(|\mathcal{F}|-|\mathcal{S}|-1)!}/_{|\mathcal{F}|!} \quad (16)$$

Finally, let $\mathrm{Sc}: \mathcal{F} \rightarrow \mathbb{R}$, that is, the SHAP score for feature i, be defined by,

Figure 3. Computation of SHAP scores for the example DT and instance $((1,1,2),\alpha)$. For each feature i, the sets to consider are all the sets that do not include the feature. 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.

$\mathcal{S}$	$\varphi(\mathcal{S})$	$\varphi(\mathcal{S}\cup\{1\})$	$\Delta(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S})\times\Delta(\mathcal{S})$
$\emptyset$	$(\Sigma_{j=1}^{6}\sigma_j)/12 + \alpha/2$	α	$\alpha/2 - (\Sigma_{j=1}^{6}\sigma_j)/12$	${}^{0!(3-0-1)!}/_{3!} = {}^{1}/_{3}$	$\alpha/6 - (\Sigma_{j=1}^{6}\sigma_j)/36$
{2}	$(\sigma_4+\sigma_5+\sigma_6)/6 + \alpha/2$	α	$\alpha/2 - (\sigma_4+\sigma_5+\sigma_6)/6$	${}^{1!(3-1-1)!}/_{3!} = {}^{1}/_{6}$	$\alpha/12 - (\sigma_4+\sigma_5+\sigma_6)/36$
{3}	$(\sigma_3+\sigma_6)/4 + \alpha/2$	α	$\alpha/2 - (\sigma_3+\sigma_6)/4$	${}^{1!(3-1-1)!}/_{3!} = {}^{1}/_{6}$	$\alpha/12 - (\sigma_3+\sigma_6)/24$
{2, 3}	$\sigma_6/2 + \alpha/2$	α	$\alpha/2 - \sigma_6/2$	${}^{2!(3-2-1)!}/_{3!} = {}^{1}/_{3}$	$\alpha/6 - \sigma_6/6$
SHAP score for feature 1					$\mathrm{Sc}(1) = \alpha/2 - (2\sigma_1 + 2\sigma_2 + 5\sigma_3 + 4\sigma_4 + 4\sigma_5 + 19\sigma_6)/72$

$\mathcal{S}$	$\varphi(\mathcal{S})$	$\varphi(\mathcal{S}\cup\{2\})$	$\Delta(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S})\times\Delta(\mathcal{S})$
$\emptyset$	$(\Sigma_{j=1}^{6}\sigma_j)/12 + \alpha/2$	$(\sigma_4+\sigma_5+\sigma_6)/6 + \alpha/2$	$-(\sigma_1+\sigma_2+\sigma_3)/12 + (\sigma_4+\sigma_5+\sigma_6)/12$	${}^{0!(3-0-1)!}/_{3!} = {}^{1}/_{3}$	$-(\sigma_1+\sigma_2+\sigma_3)/36 + (\sigma_4+\sigma_5+\sigma_6)/36$
{2}	α	α	0	${}^{1!(3-1-1)!}/_{3!} = {}^{1}/_{6}$	0
{3}	$(\sigma_3+\sigma_6)/4 + \alpha/2$	$\sigma_6/2 + \alpha/2$	$-\sigma_3/4 + \sigma_6/4$	${}^{1!(3-1-1)!}/_{3!} = {}^{1}/_{6}$	$-\sigma_3/24 + \sigma_6/24$
{1, 3}	α	α	0	${}^{2!(3-2-1)!}/_{3!} = {}^{1}/_{3}$	0
SHAP score for feature 2					$\mathrm{Sc}(2) = (-2\sigma_1 - 2\sigma_2 - 5\sigma_3 + 2\sigma_4 + 2\sigma_5 + 5\sigma_6)/72$

$\mathcal{S}$	$\varphi(\mathcal{S})$	$\varphi(\mathcal{S}\cup\{2\})$	$\Delta(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S})\times\Delta(\mathcal{S})$
$\emptyset$	$(\Sigma_{j=1}^{6}\sigma_j)/12 + \alpha/2$	$(\sigma_3+\sigma_6)/4 + \alpha/2$	$-(\sigma_1+\sigma_2+\sigma_4+\sigma_5)/12 + (\sigma_3+\sigma_6)/6$	${}^{0!(3-0-1)!}/_{3!} = {}^{1}/_{3}$	$-(\sigma_1+\sigma_2+\sigma_4+\sigma_5)/36 + (\sigma_3+\sigma_6)/18$
{2}	α	α	0	${}^{1!(3-1-1)!}/_{3!} = {}^{1}/_{6}$	0
{3}	$(\sigma_4+\sigma_5+\sigma_6)/6 + \alpha/2$	$\sigma_6/2 + \alpha/2$	$-(\sigma_4+\sigma_5)/6 + \sigma_6/3$	${}^{1!(3-1-1)!}/_{3!} = {}^{1}/_{6}$	$-(\sigma_4+\sigma_5)/36 + \sigma_6/18$
{1, 2}	α	α	0	${}^{2!(3-2-1)!}/_{3!} = {}^{1}/_{3}$	0
SHAP score for feature 3					$\mathrm{Sc}(3) = (-\sigma_1 - \sigma_2 + 2\sigma_3 - 2\sigma_4 - 2\sigma_5 + 4\sigma_6)/36$

$$\mathrm{Sc}(i;\mathcal{M},\mathbf{v}) = \sum_{\mathcal{S}\subseteq(\mathcal{F}\setminus\{i\})} \varsigma(\mathcal{S};\mathcal{M},\mathbf{v}) \times \Delta(i,\mathcal{S};\mathcal{M},\mathbf{v}) \tag{17}$$

Given an instance $(\mathbf{v}, c)$, the SHAP score 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.[37]

Misleading SHAP scores. We now demonstrate that there is surprising flexibility in choosing the influence of each feature. Concretely, we want to show that SHAP scores can yield misleading information, and that this is easy to attain. To achieve this goal, we are going to create an evidently disturbing scenario. We are going to assign values to the parameters (that is, classes) of the classifier such that feature 1 will be deemed to have *no* importance on the prediction, and features 2 and 3 will be deemed to have *some* importance on the prediction. (Such choice of SHAP scores can only be considered misleading; as argued in Figure 1d, for either κ_1 or κ_2, predicting class 1 or predicting a class other than 1 depends exclusively on feature 1.) To obtain such a choice of SHAP scores, we must have (see Figure 3),

$$Sc(1) = \alpha/2 - \left(2\sigma_1 + 2\sigma_2 + 5\sigma_3 + 4\sigma_4 + 4\sigma_5 + 19\sigma_6\right)/72 = 0 \tag{18}$$

$$Sc(2) = \left(-2\sigma_1 - 2\sigma_2 - 5\sigma_3 + 2\sigma_4 + 2\sigma_5 + 5\sigma_6\right)/72 \neq 0 \tag{19}$$

$$Sc(3) = \left(-\sigma_1 - \sigma_2 + 2\sigma_3 - 2\sigma_4 - 2\sigma_5 + 4\sigma_6\right)/36 \neq 0 \tag{20}$$

Clearly, there are arbitrarily many assignments to the values of α and $\sigma_j, j = 1,\ldots,6$ such that constraints (18), (19), and (20) are satisfied. Indeed, we can express α in terms of the other parameters $\sigma_1\ldots,\sigma_6$, as follows:

$$\alpha = \left(2\sigma_1 + 2\sigma_2 + 5\sigma_3 + 4\sigma_4 + 4\sigma_5 + 19\sigma_6\right)/36 \tag{21}$$

In addition, we can identify values for $\sigma_j, j = 1,\ldots,6$, such that the two remaining conditions (19) and (20) are satisfied. Moreover, we can reproduce the values in the DTs shown in Figure 1. Concretely, we pick $\alpha = 1 \land \sigma_1 = \sigma_3 = \sigma_4 = \sigma_6 = 0$ for both κ_1 and κ_2. Also, for κ_1, we choose $\sigma_2 = 4 \land \sigma_5 = 7$. Finally, for κ_2, we choose $\sigma_2 = 2 \land \sigma_5 = 3$. The resulting sets of SHAP scores, among others, are shown in Table 4.

The results derived above confirm that, for the two running examples, in one case the SHAP scores give a rank of the features that *must* be misleading (that is, Figure 1a), whereas for the other case, the rank of the features is less problematic, in that the most important feature is attributed the highest relative importance (that is, Figure 1c). Evidently, it is still unsettling that for the second DT, the SHAP scores of irrelevant features are non-zero, denoting some importance to the prediction that other ways of analyzing the classifier do not reveal. As the two example DTs demonstrate, and even though their operation is exactly the same when $x_1 = 1$, the predictions made for points in feature space, that are essentially *unrelated* with the given instance, can have a critical impact on the relative order of feature importance. Whereas formal explanations and adversarial examples have a precise logic-based formalization, for the two example DTs (and also for the boolean classifiers discussed later in this article), there is no apparent logical justification for the relative feature importance obtained with SHAP scores.

We refer to the misleading information produced by SHAP scores as an example of an *issue* that can be formalized as follows:

$$\forall(i \in \mathcal{F}).\quad ([\mathsf{Relevant}(i) \land (\mathrm{Sc}(i) = 0)] \lor [\mathsf{Irrelevant}(i) \land (\mathrm{Sc}(i) \neq 0)]) \tag{22}$$

The logic statement above may appear as rather specific, and so difficult to satisfy, because *every* feature must either be relevant and be assigned a zero SHAP score, or otherwise it must be irrelevant and be assigned a non-zero SHAP score. However, the example DT in Figure 1a shows that there are very simple classifiers with instances that satisfy the logic statement. Clearly, many other similarly simple examples could be devised. As we will clarify, Equation (22) is referred to as issue I8.

Revisiting the hypothetical scenario. We consider again the instance $((1,1,2),1)$, that is, an honors student from an urban household and majoring in sciences will enroll for 1 extra-

Table 4. Computed SHAP scores for the instance ((1, 1, 2), 1), by setting $\alpha = 1$ and $\sigma_1 = \sigma_3 = \sigma_4 = \sigma_6 = 0$, that is, DTs similar to those in Figures 1a and 1c, and where only the classes predicted in the terminal nodes 6 and 7 differ.

(σ_2, σ_5)	Sc(1)	Sc(2)	Sc(3)	Rank	Obs.
(4, 7)	0.000	0.083	−0.500	3,2,1	κ_1, DT in fig. 1a
(2, 3)	0.278	0.028	−0.222	1,3,2	κ_2, DT in fig. 1c
(2, 2)	0.333	0.000	−0.167	1,3,2	Same rank as for κ_2
(10, 4)	0.000	−0.167	−0.500	3,2,1	Same rank as κ_1

Table 5. Identified potential issues with SHAP scores. ($I5_{\geq}$ denotes I5 with > replaced by ≥.)

Issue	Condition	*Ik* implies *Il*?
I1	$\exists(i \in \mathcal{F}).[\mathsf{Irrelevant}(i) \land (\mathrm{Sc}(i) \neq 0)]$	
I2	$\exists(i_1, i_2 \in \mathcal{F}).[\mathsf{Irrelevant}(i_1) \land \mathsf{Relevant}(i_2) \land \lvert\mathrm{Sc}(i_1)\rvert > \lvert\mathrm{Sc}(i_2)\rvert)]$	I2 ⇒ I1
I3	$\exists(i \in \mathcal{F}).[\mathsf{Relevant}(i) \land (\mathrm{Sc}(i) = 0)]$	
I4	$\exists(i_1, i_2 \in \mathcal{F}).[\mathsf{Irrelevant}(i_1) \land (\mathrm{Sc}(i) \neq 0)] \land [\mathsf{Relevant}(i_2) \land (\mathrm{Sc}(i_2) = 0)]$	I4 ⇒ I1, I2, I3
I5	$\exists(i \in \mathcal{F}).[\mathsf{Irrelevant}(i) \land \forall(1 \leq j \leq m, j \neq i).\lvert\mathrm{Sc}(i)\rvert > \lvert\mathrm{Sc}(j)\rvert]$	I5 ⇒ I1, I2
I6	$\exists(i_1, i_2 \in \mathcal{F}).[\mathsf{Irrelevant}(i_1) \land \mathsf{Relevant}(i_2) \land (\mathrm{Sc}(i_1) \times \mathrm{Sc}(i_2) > 0)]$	I6 ⇒ I1
I7	$\exists(i_1, i_2 \in \mathcal{F}).[\mathsf{Irrelevant}(i_1) \land \mathsf{Relevant}(i_2) \land (\lvert\mathrm{Sc}(i_1)\rvert > \lvert\mathrm{Sc}(i_2)\rvert) \land (\mathrm{Sc}(i_1) > \mathrm{Sc}(i_2) > 0)]$	I7 ⇒ I1, I2, I6
I8	$\forall(i \in \mathcal{F}).([\mathsf{Relevant}(i) \land (\mathrm{Sc}(i) = 0)] \lor [\mathsf{Irrelevant}(i) \land (\mathrm{Sc}(i) \neq 0)]$	I8 ⇒ I1, I2, I3, I4, $I5_{\geq}$

curricular activity, for both DTs. Given the DTs or the TRs, formal explanations would deem the fact that she is an honors student as the answer to why the prediction is one extra-curricular activity. Also, to change the prediction, we would have to consider a non-honors student. Finally, to predict something other than one activity, while minimizing changes, we would once again have to consider a non-honors student. Unsurprisingly, for the second DT (see Figure 1c), SHAP scores concur with assigning the most importance to the fact that the student is an honors student, even though SHAP scores assign some clearly unjustified importance to the other features. More unsettling, for the first DT (see Figure 1a), SHAP scores assign importance to the fact that the student is majoring in science, that she hails from an urban household, but entirely ignore the fact that she is an honors student. And the differences between the two DTs are only the number of extra-curricular activities of non-honors students who major in arts. The reason for this abrupt change in relative feature importance (that is, 1,3,2 to 3,2,1) is inscrutable.

Table 6. Results over all 4-variable boolean functions. The two constant functions were discarded, since κ is required not to be constant. Issue I8 is not reported, because there are no occurrences.

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

More Issues with SHAP Scores

By automating the analysis of boolean functions,[13] we identified a number of issues with SHAP scores for explainability, all of which illustrate how SHAP scores can provide misleading information about the relative importance of features. The choice of boolean functions as classifiers is justified by the computational overhead involved in enumerating functions and their ranges. The list of possible issues is summarized in Table 5. Issue I8 was discussed above and its occurrence implies the occurrence of several other issues. Our goal is to uncover some of the problems that the use of SHAP scores for explainability can induce, and so different issues aim to highlight such problems.

Table 6 summarizes the percentage of functions exhibiting the identified issues, and it is obtained by analyzing *all* of the possible boolean functions defined on four variables. 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 SHAP scores. The algorithms used are detailed in earlier work,[13] and all run in polynomial-time on the size of the truth table. For example, whereas issue I5 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. Also, the two constant functions were not accounted for. Moreover, it is the case that *no* boolean classifier with four features exhibits issue I8. However, as the earlier sections highlight, there are discrete classifiers, with three features of which only one is non-boolean, that exhibit issue I8. The point here is that one should expect similar unsettling issues as the domains and ranges of classifiers increase. Besides the results summarized in this section, additional results[14] include proving the existence of arbitrary many boolean classifiers for which issues I1 to I7 occur. Furthermore, more recent work demonstrates the occurrences of some of these issues in practical classifiers, including decision trees and decision graphs represented with ordered multi-valued decision diagrams.[15]

Why do SHAP scores mislead? We identified two reasons that justify why SHAP scores can provide misleading information: the contributions of all possible subsets of fixed features are considered; and class values are explicitly accounted for. For the examples studied in this article, these two reasons suffice to *force* SHAP scores to provide misleading information.

Discussion

This article demonstrates that there exist very simple classifiers for which SHAP scores produce misleading information about relative feature importance, that is, features bearing *some* influence on a predicted class (or even determining the class) can be assigned a SHAP score of 0, and features bearing *no* influence on a predicted class can be assigned non-zero SHAP scores. In such situations there is a clear disagreement between the stated meaning of feature importance ascribed to SHAP scores[23,37,38] and the actual influence of features in predictions. It is plain that in such situations, human decision makers may assign prediction importance to irrelevant features and may overlook critical features. The results in this article are further supported by a number of recent results, obtained on arbitrary many classifiers,[13,14] that further confirm the existence of different issues with SHAP scores, all of which can serve to mislead human decision makers.

Furthermore, the situation in practice is bleaker than what these statements might suggest. For classifiers with larger number of features, recent experimental evidence suggests close to *no* correlation between the measures of feature importance obtained with tools like SHAP and those determined by rigorous computation of SHAP scores.[13] It would be rather unexpected that tools that fail to ap-

proximate what they aim to would in turn obtain more accurate measures of feature importance. A final practical remark of the results in this article is that the continued practical use of tools that approximate SHAP scores should be a reason of concern in high-risk and safety-critical domains. Recent work proposes alternatives to the use of SHAP scores.[13,40]

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, O. Létoffé, R. Passos, A. Morgado, J. Planes and N. Asher. Joao Marques-Silva also acknowledges the extra incentive provided by the ERC in not funding this research. C

The continued practical use of tools that approximate SHAP scores should be a reason of concern in high-risk and safety-critical domains.

References

1. Arenas, M., Barceló, P., Bertossi, L.E., and Monet, M. The tractability of SHAP-score-based explanations for classification over deterministic and decomposable boolean circuits. In *AAAI*, (2021), 6670–6678.
2. Arenas, M. et al. On the complexity of SHAP-score-based explanations: Tractability via knowledge compilation and non-approximability results. *J. Mach. Learn. Res. 24*, (2023), 63:1–63:58; http://jmlr.org/papers/v24/21-0389.html
3. Bach, S. et al. On pixel-wise explanations for nonlinear classifier decisions by layer-wise relevance propagation. *PloS one 10*, 7 (2015), e0130140.
4. Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M.K. Occam's razor. *Inf. Process. Lett. 24*, 6 (1987), 377–380; 10.1016/0020-0190(87)90114-1
5. Chen, H., Covert, I.C., Lundberg, S.M., and Lee, S. Algorithms to estimate shapley value feature attributions. *CoRR abs/2207.07605*, (2022); 10.48550/arXiv.2207.07605 arXiv:2207.07605
6. Chen, J., Song, L., Wainwright, M.J., and Jordan, M.I. L-shapley and C-shapley: Efficient model interpretation for structured data. In *ICLR*.
7. Datta, A., Sen, S., and Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In IEEE *S&P*, 2016, 598–617.
8. van den Broeck, G., Lykov, A., Schleich, M., and Suciu, D. On the tractability of SHAP explanations. In *AAAI*, (2021), 6505–6513.
9. van den Broeck, G., Lykov, A., Schleich, M., and Suciu, D. On the tractability of SHAP explanations. *J. Artif. Intell. Res. 74*, (2022), 851–886; 10.1613/jair.1.13283
10. Eiter, T. and Gottlob, G. The complexity of logic-based abduction. *J. ACM 42*, 1 (1995), 3–42; 10.1145/200836.200838
11. Goodfellow, I.J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *ICLR*, (2015).
12. Huang, X. and Marques-Silva, J. From Robustness to explainability and back again. *CoRR abs/2306.03048*, (2023); 10.48550/arXiv.2306.03048 arXiv:2306.03048
13. Huang, X. and Marques-Silva, J. The inadequacy of shapley values for explainability. *CoRR abs/2302.08160* (2023b); 10.48550/arXiv.2302.08160 arXiv:2302.08160
14. Huang, X. and Marques-Silva, J. A refutation of shapley values for explainability. *CoRR abs/2309.03041*, 2023; 10.48550/arXiv.2309.03041 arXiv:2309.03041
15. Huang, X. and Marques-Silva, J. Refutation of shapley values for XAI – additional evidence. *CoRR abs/2310.00416* 2023d; 10.48550/arXiv.2310.00416 arXiv:2310.00416
16. Ignatiev, A. Towards trustable explainable AI. In *IJCAI*, (2020), 5154–5158.
17. Ignatiev, A., Narodytska, N., Asher, N., and Marques-Silva, J. From contrastive to abductive explanations and back again. *AIxIA*, (2020), 335–355.
18. Ignatiev, A., Narodytska, N., and Marques-Silva, J. Abduction-based explanations for machine learning models. In *AAAI*, 2019a, 1511–1519.
19. Ignatiev, A., Narodytska, N., and Marques-Silva, J. On relating explanations and adversarial examples. In *NeurIPS*, 2019b, 15857–15867.
20. Kumar, E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S.A. Problems with shapley-value-based explanations as feature importance measures. *ICML*, (2020), 5491–5500.
21. Lipovetsky, S. and Conklin, M. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry 17*, 4 (2001), 319–330.
22. Lundberg, S.M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell. 2*, 1 (2020), 56–67; 10.1038/s42256-019-0138-9
23. Lundberg, S.M. and Lee, S. A unified approach to interpreting model predictions. *NeurIPS*, (2017), 4765–4774.
24. Marques-Silva, J. Logic-based explainability in machine learning. *Reasoning Web*, (2022), 24–104.
25. Marques-Silva, J. Disproving XAI myths with formal methods – initial results. In *ICECCS*, (2023).
26. Marques-Silva, J. and Ignatiev, A. Delivering trustworthy AI through formal XAI. *AAAI*, (2022), 12342–12350.
27. Merrick, L. and Taly, A. The explanation game: Explaining machine learning models using shapley values. *CDMAKE*, (2020), 17–38.
28. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell. 267*, (2019), 1–38.
29. Reiter, R. A theory of diagnosis from first principles. *Artif. Intell. 32*, 1 (1987), 57–95; 10.1016/0004-3702(87)90062-2
30. Ribeiro, M.T., Singh, S., and Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. *KDD*, (2016), 1135–1144.
31. Ribeiro, M.T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. *AAAI*, (2018), 1527–1535.
32. Roth, A.E. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, (1988).
33. Samek , W., et al. (Eds.). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, (2019); 10.1007/978-3-030-28954-6
34. Shapley, L.S. A value for n-person games. *Contributions to the Theory of Games 2*, 28 (1953), 307–317.
35. Shih, A., Choi, A., and Darwiche, A. A symbolic approach to explaining bayesian network classifiers. *IJCAI*, (2018), 5103–5111.
36. Slack, D., Hilgard, A., Singh, S., and Lakkaraju, H. Reliable post hoc explanations: Modeling uncertainty in explainability. *NeurIPS*, (2021), 9391–9404.
37. Strumbelj, E. and Kononenko, I. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res. 11*, (2010), 1–18; 10.5555/1756006.1756007
38. Strumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst. 41*, 3 (2014), 647–665; 10.1007/s10115-013-0679-x
39. Watson, D.S. Rational shapley values. *FAccT*, (2022), 1083–1094.
40. Yu, J., Ignatiev, A., and Stuckey, P.J. On formal feature attribution and its approximation. *CoRR abs/2307.03380*, (2023); 10.48550/arXiv.2307.03380 arXiv:2307.03380

Joao Marques-Silva was with IRIT, CNRS in Toulouse, France. He is now research professor at ICREA, University of Lleida, Spain.

Xuanxiang Huang was with the University of Toulouse, France. He is now a postdoctoral researcher with CNRS@CREATE, Singapore.