Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability

Christopher Frye[1], Ilya Feige[1], and Colin Rowat[2]

[1] *Faculty, 54 Welbeck Street, London, W1G 9XS, UK*
[2] *Department of Economics, University of Birmingham, Edgbaston B15 2TT, UK*

Abstract

Explaining AI systems is fundamental both to the development of high performing models and to the trust placed in them by their users. A general framework for explaining any AI model is provided by the Shapley values that attribute the prediction output to the various model inputs ("features") in a principled and model-agnostic way. The outstanding strength of Shapley values is their combined generality and rigorous foundation: they can be used to explain any AI system, and one always understands their values as the unique attribution method satisfying a set of mathematical axioms. However, as a framework, Shapley values are too restrictive in one significant regard: they ignore all causal structure in the data. We introduce a less-restrictive framework for model-agnostic explainability: "Asymmetric" Shapley values. Asymmetric Shapley values (ASVs) are rigorously founded on a set of axioms, applicable to any AI system, and can flexibly incorporate any causal knowledge known a-priori to be respected by the data. We show through explicit, realistic examples that the ASV framework can be used to (i) improve model explanations by incorporating causal information, (ii) provide an unambiguous test for unfair discrimination based on simple policy articulations, (iii) enable sequentially incremental explanations in time-series models, and (iv) support feature-selection studies without the need for model retraining.

1 Introduction

AI is one of the transformational technologies of our time. It has the potential to significantly improve economic productivity as well as the potential to cause widespread harm to humanity. Unfortunately, the goals of developing AI capabilities and of ensuring AI safety are not generally aligned. Helpfully, in the domain of AI explainability, this is not the case. Not only does explainability lie at the heart of AI safety, but it is also critical to the iterative development of new AI systems by exposing how they work, why they fail, and thus how they can be improved.

AI explainability can be approached in many ways. One safe starting point is to restrict one's use to *interpretable* models that require no additional explanation (e.g. linear and rules-based models). This approach is argued for in [1] in particularly sensitive settings. To explain more complex models, *model-specific* techniques leverage attributes unique to the model type in order to explain the predictions. Examples include split count for tree-based models [2] or DeepLIFT for neural networks [3]. However, model-specific approaches do not solve the problem of AI explainability in general, and, due to their bespoke nature, require that a different explanation technique be used for each model.

Model-agnostic methods provide a general approach to AI explainability that is helpful, not only for its widespread applicability, but also because of the common language it provides for AI explanations across model types. For example, permutation feature importance measures the reduction in a model's accuracy when a given feature is permuted in the data [4–6], a measure that can be meaningfully compared for models of different types. Permutation feature importance serves as a *global* explanation of what use a model makes of a given feature as it outputs predictions on an entire data set. Separate methods exist to serve a *local* explanation of an individual prediction made by the model on a specific data point [7, 8].

A local model-agnostic approach to AI explainability based on *Shapley values* is highly compelling due to its principled mathematical foundation [9] and its ability to capture all the interactions between features that lead to a model's prediction. Shapley values have been used in AI explainability for decades [10–13], with the general framework articulated more recently in [14].

Despite their strength relative to other model-agnostic techniques, the Shapley-value approach has four outstanding shortcomings: (a) they are computationally expensive, (b) like many other techniques, they rely on unrealistic fictitious data, (c) they ignore causality, and (d) they provide explanations based on the raw model-input features, which may themselves not be directly amenable to semantic interpretation. The computational cost, (a), can be reduced, either by Monte Carlo sampling, or by more efficient model-specific estimation techniques as in [15]. A solution to the fictitious data problem (b) is presented in the forthcoming work of [16], summarised here in Sec. 2.3 and App. A. This paper developsthe first approach, to our knowledge, to incorporating causality (c) into the Shapley framework.

Addressing causality in model explainability should not be considered optional. Causality is at the heart of understanding any system, and this is no different for an AI system. However, causal deduction is difficult. Indeed, one of the paradigmatic advantages of modern machine learning is its ability to extract highly predictive correlations from large data sets utilising high-capacity models and efficient learning algorithms, instead of focusing on potentially simpler causal understandings.

The field of *causal inference* does provide a rigorous framework for understanding causality, given a causal graph and somewhat restrictive assumptions [17–20]. However, the problem of ascertaining the causal graph remains difficult. *Causal discovery* methods exist to automatically extract causal graphs, but performance across different methods is highly variable [21]. Moreover, in machine learning, it is exceedingly rare that the full causal model underlying the data is known, since data sets often contain hundreds to thousands of features. Therefore, explainability methods should be able to incorporate known causal relationships without the prohibitive requirement of full knowledge of the causal graph.

In this work, we generalise the Shapley-value framework to incorporate causal knowledge. Critically, we do this in a way that preserves the axiomatic mathematical construction of the framework, while introducing the opportunity to handle any amount of causal knowledge. In particular, the framework does not require specification of the complete causal graph underlying a model's data. An equivalence class of causal graphs could serve as a natural starting point, and our approach can lead to useful insights even when just a small fraction of causal relationships are known.

The incorporation of causal knowledge into AI explainability is not merely an academic pursuit. There are important societal concerns surrounding AI, the resolution of which demand model explanations built upon a rigorous causal foundation. Among these is the concern that consequential AI models propagate unfair bias and discrimination. While practical methods are available to impose statistical fairness on a model's decisions in the aggregate [22–26], much more machinery (e.g. metrics or full causal models) is generally required to ensure a model makes fair decisions for individuals, consistent with the underlying causality [27–30]. The framework developed in this paper bridges this gap by introducing a simple, practical approach to measuring causal fairness.

In total, our main contributions in this work are threefold:

1. In Sec. 2, we clarify and expand upon the Shapley-value framework for model-agnostic explainability. We introduce a definition for *global Shapley values* that naturally sum to a variant of the model's accuracy. We further show how to modify the value function underlying the framework to produce model explanations that respect correlations in the data.

2. As our primary contribution, in Sec. 3 we introduce the formal theoretical framework that allows one to incorporate causal knowledge into model-agnostic explainability. We do so by relaxing 1 of the 4 axioms (symmetry) underpinning the Shapley-value framework. We present *Asymmetric Shapley values*, as the more flexible approach to explainability capable of incorporating causal information into a better understanding of the model's behaviour.

3. In Sec. 4, we present four applications of the ASV framework: (i) incorporating partial causal understanding of a data set into the explanation of its model; (ii) a generic test of unresolved unfair discrimination [28] that can realistically be applied in practical applications; (iii) sequential feature importance in time-series modelling; and (iv) support for feature-selection studies by predicting model accuracies achievable on subsets of the data's features.

2 Shapley values for model-agnostic explainability

In cooperative game theory [31] a team of players $N = \{1, 2, \ldots, n\}$ work together to earn a certain amount of value $v(N)$. Here v is a value function $v : 2^N \to \mathbb{R}$ that associates a real value $v(S)$ with any coalition $S \subseteq N$. Since the value function indicates how effective each coalition of players is on its own, it can be used to decide how credit for the total earnings $v(N)$ should be distributed among the individual players. The *Shapley value* $\phi_v(i)$ offers one well-motivated approach to attribution among the players $i \in N$:

$$\phi_v(i) = \sum_{S \subseteq N \setminus i} \frac{s!\,(n-s-1)!}{n!}\,\left[v(S \cup i) - v(S)\right] \tag{1}$$

where $s = |S|$ and i is used to denote the singleton $\{i\}$ for simplicity. Shapley values uniquely satisfy a natural set of axioms that will be detailed in Sec. 2.1 below [9].

The combinatorial weight in this average over coalitions S serves to make this definition of Shapley value $\phi_v(i)$ equivalent to a more intuitive average over permutations in the following way. Let Π denote the set of all permutations of N. For an individual permutation $\pi \in \Pi$, the inequality $\pi(j) < \pi(i)$ means that j precedes i under ordering π. Then the Shapley value $\phi_v(i)$ can be equivalently rewritten as the following average over Π:

$$\phi_v(i) = \sum_{\pi \in \Pi} \frac{1}{n!} \Big[v(\{j : \pi(j) \leq \pi(i)\}) - v(\{j : \pi(j) < \pi(i)\})\Big] \tag{2}$$

The weight $1/n!$ makes this a uniform average over permutations, i.e. over orders in which the team can be "built up" from the empty set to N. This leads to the following interpretation: $\phi_v(i)$ is the marginal contribution that player i makes upon joining the team, averaged over all orders in which the team can be formed.

One goal for model explainability in machine learning is to compute the importance of each of a model's features in determining its output. Shapley values, with the model's features interpreted as the players in a cooperative game, offer a well-controlled approach to feature importance that has appeared in the explainability literature [10–14]. To fully define Shapley values for a given model, one need only define the value function v corresponding to that model acting on coalitions of its features.

To make this precise, let us consider a classification problem in which each data point x in data set $\mathcal{X}$ has a label $y \in \{1, 2, \ldots, k\}$ corresponding to which of k classes it falls into. A model $f : \mathcal{X} \to [0, 1]^k$ maps each data point x to a k-dimensional vector that corresponds to a categorical probability distribution over the k classes. The model is trained so that the y-component $f_y(x)$ of the model's output is its predicted probability that x belongs to class y; for consistency $\sum_{y=1}^{k} f_y(x) = 1$. We can explain the behaviour of this model using Shapley values.

To define a value function v corresponding to a model's prediction $f_y(x)$, we need to define the way that f acts on a subset x_S of x's features. Since f is only defined on the full input vector x, it is standard practice [14] to marginalise over the excluded features $x_{\bar{S}}$, where $\bar{S} = N \setminus S$, in the following way:

$$v_{f_y(x)}(S) = \mathbb{E}_{p(x')}\left[f_y(x_S \sqcup x'_{\bar{S}})\right] \tag{3}$$

The expectation is over $p(x')$, the probability distribution from which the unlabelled data set $\mathcal{X}$ is a set of samples, and $x_S \sqcup x'_{\bar{S}}$ is the spliced data point that combines in-coalition features from x with out-of-coalition features from x'. This standard definition of the value function $v_{f_y(x)}$ is problematic because it ignores correlations between features in S and $\bar{S}$, and we will modify it in Sec. 2.3.

The value function of Eq. (3) leads directly, through the average over permutations in Eq. (2), to Shapley values that explain an individual prediction $f_y(x)$ made by the model under investigation at data point x. Shapley values thus offer a method to perform local model explainability.

2.1 Axioms satisfied by Shapley values

As an attribution method, Shapley values uniquely satisfy the following four axioms [9]:

- **Axiom 1 (Efficiency)** $\sum_{i \in N} \phi_v(i) = v(N) - v(\{\})$.

  This can be understood from Eq. (2), from which it can be seen that the sum over i is telescoping. For the purpose of model explainability, this axiom states that the model's prediction $f_y(x)$ is fully distributed among the features in the data:

  $$\sum_{i \in N} \phi_{f_y(x)}(i) = f_y(x) - \mathbb{E}_{p(x')}\big[f_y(x')\big] \tag{4}$$

  The offset term $v(\{\})$, which is usually set to zero in the game theory context, is the average probability (over all data points x') that f assigns to class y. This baseline prediction, not attributable to any individual feature i, is related to the class balance in the data.

- **Axiom 2 (Linearity)** $\phi_{\alpha u + \beta v} = \alpha\, \phi_u + \beta\, \phi_v$ *for any value functions* u, v *and any* $\alpha, \beta \in \mathbb{R}$.

  In model explainability, this states that the Shapley values for a linear ensemble model are the linear combination of Shapley values for the members of the ensemble.

- **Axiom 3 (Null player)** $\phi_v(i) = 0$ *whenever* $v(S \cup i) = v(S)$ *for all* $S \subseteq N \setminus i$.

  For model explainability, this states that if a feature is completely disconnected from a model's output, it receives zero Shapley value.

- **Axiom 4 (Symmetry)** $\phi_v(i) = \phi_v(j)$ *whenever* $v(S \cup i) = v(S \cup j)$ *for all* $S \subseteq N \setminus \{i, j\}$.

  This assumption drives the uniform distribution over permutations in Eq. (2).

  For the purpose of model explainability, this axiom states that feature importance is equally distributed over features that are identically informative about the model's prediction.

The theoretical control offered by these axioms and the consequential uniqueness of Shapley values are coveted properties for many applications. However, we will see in Sec. 3.1 that often these 4 axioms are too restrictive in the application of model explainability.

2.2 Global Shapley values

While the Shapley values defined by substituting Eq. (3) into Eq. (2) provide a local explanation of the model's individual prediction $f_y(x)$, it is possible to derive a global explanation of the model's behaviour by aggregating local ones. While there are several ways one might design the aggregation procedure, we find it useful to perform an average over a labelled data set. To be explicit, we define *global Shapley values* for model f as follows:[1]

$$\phi_f(i) \;=\; \mathbb{E}_{p(x,y)}\big[\phi_{f_y(x)}(i)\big] \tag{5}$$

Here $p(x, y)$ is the probability distribution over data x and labels y that model f aims to predict (i.e. the distribution from which the labelled data are samples), and $\phi_{f_y(x)}(i)$ is the (local) Shapley value explaining the model's individual prediction $f_y(x)$ at labelled data point (x, y). We omit the subscript $y(x)$ in the global Shapley value to indicate that it is evaluated over the whole data set, not just at x.

With this definition, global Shapley value $\phi_f(i)$ can be interpreted as the portion of model f's accuracy attributable to feature i. This interpretation follows from the fact that global Shapley values satisfy

$$\sum_{i \in N} \phi_f(i) \;=\; \sum_{i \in N} \mathbb{E}_{p(x,y)}\big[\phi_{f_y(x)}(i)\big] \;=\; \mathbb{E}_{p(x,y)}\big[f_y(x)\big] - \mathbb{E}_{p(x')}\mathbb{E}_{p(y)}\big[f_y(x')\big] \tag{6}$$

using Eq. (5) and Eq. (4), respectively. The first term on the right can be interpreted as model f's accuracy. More precisely, it is the accuracy one achieves by randomly drawing predicted labels according to f's predicted categorical distribution.[2] The second term is an offset corresponding to the accuracy one would be left with if the prediction for data point x was made using the model's output $f(x')$ on a

[1] Global Shapley values can either be viewed as an aggregation of local Shapley values, as in Eq. (5), or more directly as the Shapley values that follow from a global value function, Eq. (7).

[2] This is similar to, but not the same as, the accuracy one achieves by predicting $\text{argmax}_y f_y(x)$ for x's label.

randomly drawn data point x'. This baseline accuracy, not attributable to any individual feature i, is related to the class balance of the data.

In what follows, it will be helpful to define simplifying notation for the *global value function*:

$$\mathcal{A}_f(S) = \mathbb{E}_{p(x,y)}\left[v_{f_y(x)}(S)\right] \tag{7}$$

which can be interpreted as model f's accuracy if it is required to marginalise over features excluded from S when making its predictions. Then the global sum in Eq. (6) can be rewritten as

$$\sum_{i \in N} \phi_f(i) = \mathcal{A}_f(N) - \mathcal{A}_f(\{\}) \tag{8}$$

That is, global Shapley values sum to the model's accuracy (on all its features N) minus a baseline term corresponding to the performance of the model's average output.

Global Shapley values offer a way to understand model f's behaviour over an entire data set, while remaining consistent with local explanations for the model's output on individual data points. To the best of our knowledge, this definition of global Shapley values, which is most natural due to its connection to model f's accuracy, has not appeared elsewhere in the literature.

2.3 Model explanations that respect the data manifold

Shapley values for model explainability are widely based on the value function of Eq. (3), but the way that spliced data points $x_S \sqcup x'_{\bar{S}}$ are drawn in Eq. (3) is problematic. Because x' is drawn unconditionally from the data distribution $p(x')$, its out-of-coalition features $x'_{\bar{S}}$ may not be compatible with the in-coalition features x_S of the data point under investigation. For example, in the Census Income data explored in Sec. 4.1 below, x_S could represent "marital status = never married" and $x'_{\bar{S}}$ could be drawn as "relationship = husband". We refer to such an impossible data point as lying *off the data manifold.*

Such unrealistic data is not like any of the model's training data, nor does it resemble data on which the model will be deployed. Incompatible splices lie outside the model's region of validity, and there is no reason to believe that the model will behave sensibly on such inputs. Model explanations based on unrealistic data hinder insight into the model's actual behaviour on real data. This is a flaw with most approaches to model explainability; for one exception, see [32].

To fix this problem, one should replace $p(x')$ in the off-manifold value function $v_{f_y(x)}(S)$ of Eq. (3) with the conditional distribution $p(x'|x_S)$, so that

$$u_{f_y(x)}(S) = \mathbb{E}_{p(x'|x_S)}\left[f_y(x_S \sqcup x'_{\bar{S}})\right] \tag{9}$$

We will refer to $u_{f_y(x)}(S)$ as the on-manifold value function. This quantity is nontrivial to compute, because for high-dimensional or continuous data the empirical $p(x'|x_S)$ cannot be used to estimate the integral. In [16] a method is developed to learn $p(x'|x_S)$ using variational inference. We will refer to any learnt model of $p(x'|x_S)$ as a probabilistic data imputer. The implementation we used in the experiments of Sec. 4 is developed in [16] and summarised in App. A. For an empirical approach to estimating on-manifold Shapley values, see [33].

The focus of this paper is to incorporate causal knowledge into model-agnostic explainability, and this cannot be done without first getting the correlations right. For this reason, the on-manifold value function $u_{f_y(x)}(S)$ of Eq. (9) will be assumed throughout the remainder of the paper unless explicitly indicated otherwise. (Shapley values computed off- and on-manifold will be compared in Sec. 4.1.)

3 Asymmetric Shapley values for better model explainability

In this section, we present a theoretical framework to incorporate causal knowledge about the data-generating process into model-agnostic explainability.

3.1 Argument against symmetry

In Secs. 2.1 and 2.2 we discussed the utility of Axioms 1 – 3 (Efficiency, Linearity, and Null player) satisfied by Shapley values. Axiom 4 (Symmetry) is a reasonable initial expectation for an explanation method:

it places all features on equal footing for attribution. This means that Shapley values will uniformly distribute feature importance over identically informative (i.e. redundant) features. However, when redundancies exist in the data, we might instead seek a sparser explanation of the model's behaviour. Instead of uniformly distributing feature importance over redundant features, we might instead prefer to concentrate the importance on those features we deem more fundamental. This would require us to relax Axiom 4.

Before eliminating Axiom 4, consider when the axiom is active in model explainability. To do this, suppose that the value function is symmetric: $u_{f_y(x)}(S \cup i) = u_{f_y(x)}(S \cup j)$ for all $S \subseteq N \setminus \{i, j\}$. Referring to Eq. (9) and letting $\tilde{S}$ denote $N \setminus (S \cup \{i, j\})$, this means that

$$\mathbb{E}_{p(x'|x_{S\cup i})}\big[f_y(x_{S\cup i} \sqcup x'_{\tilde{S}\cup j})\big] = \mathbb{E}_{p(x'|x_{S\cup j})}\big[f_y(x_{S\cup j} \sqcup x'_{\tilde{S}\cup i})\big] \tag{10}$$

Since $f_y(x)$ is the probability that the model f assigns to class y for data point x, let us write this momentarily as $p(\hat{y} = y|x)$, where $\hat{y}$ is the predicted class rather than the actual class in the labelled data. Then Eq. (10) becomes

$$\int dx'_{\tilde{S}\cup j}\; p(x'_{\tilde{S}\cup j}|x_{S\cup i})\; p(\hat{y} = y|x_{S\cup i} \sqcup x'_{\tilde{S}\cup j}) = \int dx'_{\tilde{S}\cup i}\; p(x'_{\tilde{S}\cup i}|x_{S\cup j})\; p(\hat{y} = y|x_{S\cup j} \sqcup x'_{\tilde{S}\cup i}) \tag{11}$$

which implies that

$$p(\hat{y} = y|x_{S\cup i}) = p(\hat{y} = y|x_{S\cup j}) \tag{12}$$

If this is to hold for all $S \subseteq N \setminus \{i, j\}$ then x_i and x_j must contain identical information for predicting the model f's output, given any other features x_S that might be known as well.

Eq. (12) could hold trivially if $f_y(x)$ is disconnected from x_i and x_j, but this case is already covered by Axiom 3 (Null player). Assuming x_i and x_j are nontrivial components of the data generating process, the more common situation[3] is that they are deterministically (bijectively) related to one another. This is exactly the case in which one might want control over which of x_i and x_j is attributed importance. For example, if x_i is known to be the deterministic causal ancestor of x_j, one might want to attribute all the importance to x_i and none to x_j. (See Section 3.3 for worked examples.) However, Axiom 4 (Symmetry) obeyed by Shapley values would require importance to be distributed evenly between x_i and x_j.

3.2 Asymmetric Shapley values

We have so far argued that symmetry should not be considered a requirement for model-agnostic explainability, since it can obfuscate known causal relationships in the data. Interestingly, relaxing this axiom still provides a rich theory that has been explored in the cooperative game theory literature.

Above Eq. (2) we defined Π as the set of all permutations on finite set $N = \{1, 2, \ldots, n\}$. Let us further define $\Delta(\Pi)$ as the set of probability measures on Π. That is, each $w \in \Delta(\Pi)$ is a map $w : \Pi \to [0, 1]$ satisfying $\sum_{\pi\in\Pi} w(\pi) = 1$. Then we define *Asymmetric Shapley values* with respect to a value function u and a distribution $w \in \Delta(\Pi)$ as

$$\phi_u^{(w)}(i) = \sum_{\pi\in\Pi} w(\pi) \Big[u(\{j : \pi(j) \leq \pi(i)\}) - u(\{j : \pi(j) < \pi(i)\})\Big] \tag{13}$$

These values are known as "quasivalues" or "random order values" in cooperative game theory [34, 35]. ASVs uniquely satisfy Axioms 1 – 3 (Efficiency, Linearity, and Null player). They do not satisfy Axiom 4 (Symmetry) unless $w \in \Delta(\Pi)$ is taken to be uniform, as in $w(\pi) = 1/n!$ in which case they reduce to the (symmetric) Shapley values of Eq. (2).

The key point for model explainability is that ASVs allow the practitioner to place a non-uniform distribution over the orderings in which features are fed to the model when determining each feature's impact on the model's prediction. The flexibility one gains in choosing $w(\pi)$ makes the ASV framework useful for a variety of applications in which a preferred ordering exists for (perhaps a subset of) a model's features. For example, it might be known that one feature is the causal ancestor of another; $w(\pi)$ could then be chosen to only place nonzero weight on permutations π that order the ancestor before its descendant. A general guideline for incorporating causal knowledge into $w(\pi)$ is given in Eq. (18). See Sec. 4 for this and several other interesting use cases.

[3] Another possibility is that x_i and x_j are i.i.d. and appear symmetrically in the structural equations determining $\hat{y}$ from x. In this case, Eq. (18) would place x_i and x_j on equal footing when computing ASVs.

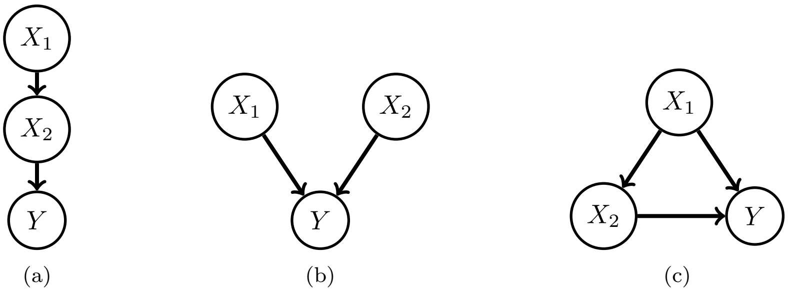


Figure 1: Example causal graphs depicting 3 possible data generating processes, in which an outcome Y is caused by features X_1 and X_2.

3.3 Examples in two dimensions

Consider the simple case of $N = \{1, 2\}$ in which a model $f(x_1, x_2)$ uses 2 features to predict a discrete outcome y. That is, $f_y(x_1, x_2)$ is trained to approximate a data distribution $p(Y = y|X_1 = x_1, X_2 = x_2)$. To explain this model's output $f_y(x_1, x_2)$ on a specific input, we define a value function as in Eq. (9):

$$\begin{aligned} u(\{\}) &= \mathbb{E}_{p(x_1',x_2')}\left[f_y(x_1', x_2')\right] \\ u(\{1\}) &= \mathbb{E}_{p(x_2'|x_1)}\left[f_y(x_1, x_2')\right] \\ u(\{2\}) &= \mathbb{E}_{p(x_1'|x_2)}\left[f_y(x_1', x_2)\right] \\ u(\{1,2\}) &= f_y(x_1, x_2) \end{aligned} \tag{14}$$

where we omit the subscript on $u_{f_y(x)}$ for clarity. Shapley values then follow from Eq. (2):

$$\begin{aligned} \phi_u(1) &= \frac{1}{2}\Big[u(\{1\}) - u(\{\})\Big] + \frac{1}{2}\Big[u(\{1,2\}) - u(\{2\})\Big] \\ \phi_u(2) &= \frac{1}{2}\Big[u(\{1,2\}) - u(\{1\})\Big] + \frac{1}{2}\Big[u(\{2\}) - u(\{\})\Big] \end{aligned} \tag{15}$$

The first bracketed term on the right in each equation corresponds to the permutation $\pi = (12)$, the second to $\pi = (21)$; Symmetry Axiom 4 sets these to $\frac{1}{2}$ each. The sum rule of Axiom 1 (Efficiency) is clearly satisfied.

While Shapley values are immediately fixed by the choice of value function in Eq. (14), ASVs provide a layer of flexibility, allowing the practitioner to choose $w(\pi)$ according to the application:

$$\begin{aligned} \phi_u^{(w)}(1) &= w_{(12)}\big[u(\{1\}) - u(\{\})\big] + w_{(21)}\big[u(\{1,2\}) - u(\{2\})\big] \\ \phi_u^{(w)}(2) &= w_{(12)}\big[u(\{1,2\}) - u(\{1\})\big] + w_{(21)}\big[u(\{2\}) - u(\{\})\big] \end{aligned} \tag{16}$$

where $w_{(12)} + w_{(21)} = 1$. This guarantees that Axiom 1 is still satisfied.

Now suppose some basic information about the causal model underlying the data is known. Three (non-exhaustive) possibilities are shown in Fig. 1. If the data is generated according to Fig. 1(a) then one might choose $w_{(12)} = 1$ and $w_{(21)} = 0$, so that

$$\begin{aligned} \phi_u^{(w)}(1) &= u(\{1\}) - u(\{\}) \\ \phi_u^{(w)}(2) &= u(\{1,2\}) - u(\{1\}) \end{aligned} \tag{17}$$

In words, $\phi_u^{(w)}(1)$ reports the impact of x_1 on f as compared to its average baseline output $u(\{\})$, while $\phi_u^{(w)}(2)$ reports the marginal impact of x_2 on the model f that has already received x_1. If instead the data is generated by Fig. 1(b) then it is natural to choose $w_{(12)} = w_{(21)} = 1/2$, so that each $\phi_u^{(w)}(i)$ reduces to the (symmetric) Shapley value $\phi_u(i)$ of Eq. (15).

In general, to incorporate any fixed amount of causal knowledge into the model explanation, a simple and natural way to link the Shapley permutation probabilities to this causal knowledge is

$$w_{\text{causal}}(\pi) = \frac{1}{\text{const}} \times \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ for all known ancestors } i \text{ of descendants } j \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

where the constant enforces normalisation. Note that this reduces to the uniform weight $w(\pi) = 1/n!$ of symmetric Shapley values if no causal information is known, since in that case the condition of Eq. (18) is vacuously satisfied. Using Eq. (18) as a guideline reduces the practitioner's freedom in designing $w(\pi)$, which was perhaps too vast to be useful. Following Eq. (18), one would set $w_{(12)} = 1$ and $w_{(21)} = 0$ in the case of Fig. 1(c) as well.

3.4 A data agnosticism continuum

Shapley values provide a maximally data-agnostic explanation of a model f by uniformly averaging over all orderings in which coalitions of features can be constructed. Thus, absolutely no information about the causal structure of the data is incorporated into an explanation based on Shapley values. At the other extreme, the goal of *causal inference* is to infer the exact causal process underlying the data. ASVs span this data-agnosticism continuum in the sense that they allow any knowledge about the data, however incomplete, to be incorporated into an explanation of the model's behaviour. For example, one might choose $w(\pi)$ so that a certain feature is always ordered first, but that orderings over the remaining features are otherwise uniformly weighted. Alternatively, one might choose to specify a single permutation in the sum of Eq. (13), corresponding perhaps to a fully specified causal graph. These extremes are explored further in the explicit examples of Secs. 4.1 and 4.3.

In this way, ASVs enable some *a priori* knowledge about the data-generating process to be incorporated into the model's explanation, while not requiring the often-prohibitive requirement of full causal inference, i.e. of full knowledge of the causal model underlying the data.

4 Applications and experimental results

In this section, we consider applications of ASVs. We demonstrate through examples that ASVs offer useful insights in the following scenarios: (i) when something is known about the causal structure underlying a model's data, (ii) when there are subtle questions about unfair discrimination sensitive to the underlying causality, (iii) when the data type under study possesses an intrinsic ordering, and (iv) when one is interested in the predictivity of a subset of the model's features.

4.1 Causality-based model explanations

First we demonstrate the insights that can be gleaned through the ASV framework when something is known about the causal structure underlying a model's data. To do this, we performed experiments using the Census Income data set from the UCI ML repository [36]. The data contains about 49k unique individuals from the US Census Bureau's 1994 and 1995 population surveys. Each record contains 13 features and a binary label indicating whether annual income exceeded \$50k. Some features in the data are clearly causal ancestors of others (e.g. "age" causally preceding "education"), thus making this a good use-case for ASVs.

We trained a neural-network classifier on the Census Income data; see App. B.1 for details. While the data has a 76/24 class balance, the classifier achieves about 85% accuracy, consistent with the accuracies reported in the UCI ML data description. This is the model f we aim to explain using ASVs.

To set a baseline, we first computed global (symmetric) Shapley values for this model. We did this by sampling the sum in Eq. (2), using the value function of Eq. (3), and aggregating as in Eq. (5). This baseline calculation does not respect correlations in the data, as discussed in Sec. 2.3. It evaluates the model on fictitious spliced data points that may lie far from the model's region of validity. For example, to evaluate the model on a coalition x_S representing "marital status = never married", this coalition is spliced together with randomly drawn data points $x'_{\bar{S}}$, 40% of which contain "relationship = husband". This baseline model explanation is labelled "Off data manifold" in Fig. 2.

Next, we computed global (symmetric) Shapley values that respect the correlations in the model's data. We computed these by sampling the sum in Eq. (2), using the value function of Eq. (9), and aggregating according to Eq. (5). The presence of $p(x'_{\bar{S}}|x_S)$ in Eq. (9) eliminates the problem of unrealistic fictitious data. For $p(x'_{\bar{S}}|x_S)$ we trained a neural-network data imputer as described in App. A. The resulting values are labelled "On data manifold" in Fig. 2.

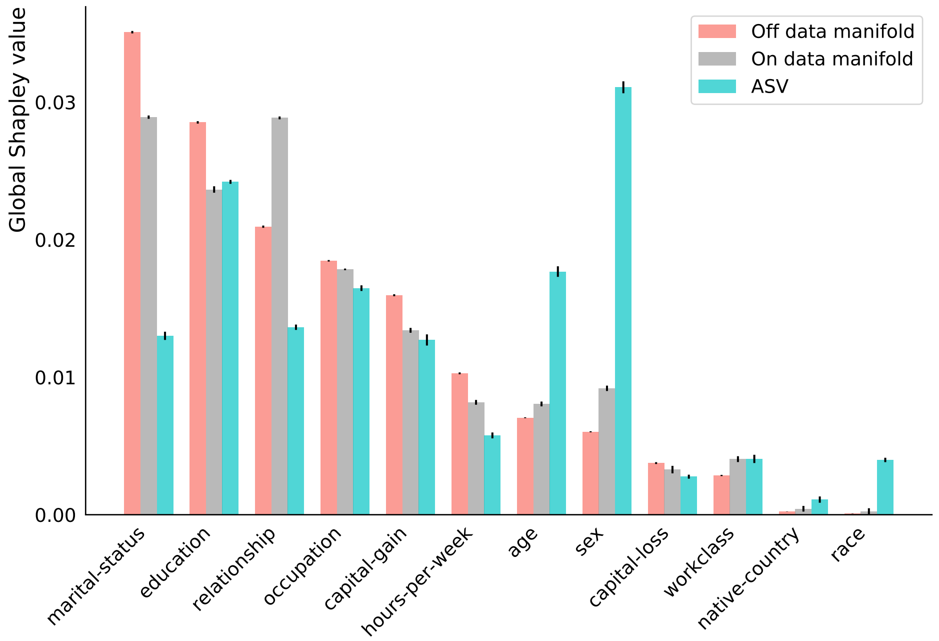


Figure 2: Global explanations of a model trained on Census Income data [36]. "Off data manifold" and "On data manifold" Shapley values place all features on equal footing in the average over coalitions, while ASVs place demographic features (decided at birth) ahead of non-demographic features (decided later) in the coalitions. Off-manifold values do not take correlations into account when choosing fictitious data for the computation, while on-manifold and asymmetric values construct realistic data from a learnt distribution. Off-manifold values represent the standard approach to Shapley explanations at present, while ASVs are the primary contribution of this paper.

Finally, we computed global ASVs for this model by sampling the sum in Eq. (13), using the value function of Eq. (9), and aggregating according to Eq. (5). We incorporated a basic causal understanding of the data into our choice of weight factors $w(\pi)$, placing nonzero weight only on those permutations in which the features determined at birth (age, sex, native country, and race) arise before the features determined later. To be explicit:

$$w_{\text{causal}}(\pi) = \frac{1}{4!\,8!} \times \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ for all } i \in A \text{ and all } j \in D \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

where

$$\begin{aligned} A &= \{\text{age, sex, native country, race}\} \\ D &= \{\text{marital status, education, relationship, occupation,} \\ &\qquad \text{capital gain, hours per week, capital loss, work class}\} \end{aligned} \tag{20}$$

This follows the guideline set in Eq. (18) since each known causal ancestor (in A) is required to appear before its known causal descendants (in D). The resulting values are labelled "ASV" in Fig. 2. The ASVs for features in A correspond to the model accuracy attributable to these features, assuming set D is yet unknown. The ASVs of features in D indicate the additional model accuracy gained from these features, assuming set A is already known.

Several interesting trends are apparent in Fig. 2. Let's focus first on the (symmetric) Shapley values. Off-manifold values give insight into the model f's functional form, regardless of any correlations in the data. For example, the model exhibits strong direct dependence on marital status, more so than it does on relationship. On-manifold values describe how the model f's output varies with each feature on the data manifold, thus taking correlations in the data into account. For example, marital status and relationship are so tightly correlated that they give rise to equal on-manifold Shapley values. ASVs, on the other hand, are computed on the data manifold but place certain features (here, in A) ahead of others (in D) when attributing importance. This results in the ASVs of features in A being greater than or equal to the corresponding on-manifold Shapley values, with strict inequality only when an upstream feature is predictive of a downstream feature that the model uses to make its predictions. One helpful constraint on the 3 explanations of Fig. 2 is that they all satisfy the same sum rule, Eq. (6).

Most interestingly, Fig. 2 shows that, while sex has a relatively small Shapley value, it receives the largest ASV. This means that, in this data set, sex explains enough variation in downstream features — most visibly marital status, relationship, occupation, and hours per week — to be a very strong predictor of annual income on its own. Quantitatively, recall that the model f's accuracy is 85%, and that the 76/24 class balance prevents us from attributing the majority of this accuracy to any individual feature. Sex's ASV indicates that roughly 3% of the accuracy[4] can be attributed to sex, which is quite significant given the class imbalance. This example demonstrates that meaningful insights can be extracted with only a crude causal understanding of the data.

4.2 Causal explanations of unfair discrimination

Unfair discrimination in machine learning is a difficult problem to solve. It is insufficient to merely withhold sensitive information like gender or race from a model, since these attributes correlate uncontrollably with many other features. Methods exist to impose constraints, such as demographic parity, on the decisions of a model in aggregate [22–26], but many alternative definitions of statistical fairness exist [37], and, worse, they cannot be simultaneously satisfied [38, 39]. More satisfying definitions of fairness that focus on individuals and underlying causal processes [27–30] require much more machinery to measure or impose, often prohibitively so. See [40] for an introduction to this field. In this section, we demonstrate that the ASV framework can supply a practical measure of causal unfairness.

To explore causal unfairness in a controlled setting, we performed experiments on two synthetic college-admissions data sets. We took inspiration from the Berkeley admissions controversy of 1973 in designing this synthetic data [41]. Each data set has 3 features:

$$X_1 = \text{gender}, \quad X_2 = \text{test score}, \quad X_3 = \text{department choice}, \tag{21}$$

and a binary label Y = college admission. We took X_1 and X_3 to be binary and X_2 to be continuous. We will refer to the two synthetic data sets, which will be described in detail below, as "fair" and "unfair".[5]

We generated the first data set using the causal graph of Fig. 3(a). While more men than women are admitted to university in this "fair" data set — 62% versus 38% — this only occurs because a larger fraction of women applied to the more competitive department than men. See App. B.2 for the explicit structural equations of the data generating process.

We constructed the second data set using the causal graph of Fig. 3(b). In this graph, there is an additional pathway through which gender (X_1) can affect admission (Y), i.e. through

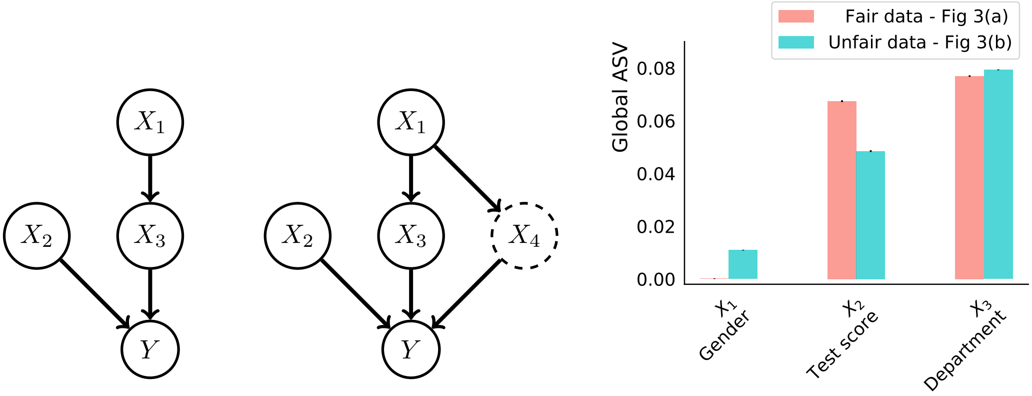
$$X_4 = \text{unreported referral}. \tag{22}$$

In this data set, men recommend other men for admission more often than other women, and for the purpose of this experiment we consider this causal pathway to be unfair. Still, this is not directly visible in the data set: only features X_1, X_2, and X_3 are reported. In this unfair data set, 64% of men and 36% of women are admitted to university, quite similar to the fair data set. Nevertheless, we will show that the ASV framework has the capacity both to verify the fairness of the first data set and expose the unfairness of the second.

To do this, we trained a neural-network classifier on each of the synthetic data sets. See App. B.2 for details about architecture and hyperparameters. While each synthetic data set has a 50/50 class balance between admissions and rejections, these classifiers achieved around 73% accuracy. Since the output of each classifier varies with all 3 recorded features, the corresponding (symmetric) Shapley values are all generically nonzero and incapable of judging whether or not unfair processes exist in the admissions process. ASVs, on the other hand, are capable of isolating these unfair processes.

[4] That is, 3 out of 85 in an absolute (not relative) sense. See Eq. (6) for the precise variant of "accuracy" relevant here.

[5] To be sure, fairness is open to ethical interpretation, and many alternative definitions exist; see e.g. [37]. The variant of fairness satisfied here is developed in [28] and summarised below.



(a) (b) (c)

Figure 3: (a) Fair causal graph underlying synthetic college-admissions data set, in which x_1 = gender, x_2 = test score, x_3 = department choice, and y = admission. (b) Unfair causal graph underlying 2nd synthetic data set, in which an additional unobserved variable, x_4 = unreported referral, influences the admissions process. (c) Global ASVs for models trained on these 2 synthetic data sets. Values for gender indicate the effect that gender, x_1, has on model predictions outside the fair pathway through department choice, x_3. The unfairness of unreported referrals, x_4, is thus exposed.

We computed global ASVs for these models by sampling the sum in Eq. (13), using the value function of Eq. (9), and aggregating according to Eq. (5). For $p(x'_{\bar{S}}|x_S)$, we trained a neural-network data imputer as described in App. A. For $w(\pi)$, we placed nonzero weight only on those permutations in which gender arises after department choice:

$$w_{\text{fair}}(\pi) = \frac{1}{3} \times \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ for all } i \in R \text{ and all } j \in S \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

where in this case

$$R = \{\text{department choice}\}$$
$$S = \{\text{gender}\}$$

The resulting ASV for gender then indicates how much additional model accuracy is gained from gender when department choice is already known. The ASVs for the 2 synthetic data sets and corresponding models are shown in Fig. 3(c). Note that this framework is able to verify the fairness of the data generated without unreported referrals, Fig. 3(a), by assigning an ASV of nearly zero (0.0002 ± 0.0001) to gender. Likewise, this approach exposes the unfairness of the data generated *with* unreported referrals, Fig. 3(b), by assigning a significant nonzero value (0.0105 ± 0.0001) to gender.

A more general study of fairness in realistic examples can easily be accommodated by the ASV framework as well. One must simply define w_{fair} to require that all *resolving variables* (R) appear before all *sensitive attributes* (S). Here we use the language of causal fairness developed by [28], where a resolving variable is any variable whose influence by sensitive attributes is deemed fair.[6] ASVs defined using w_{fair} thus generally check whether sensitive attributes affect a model's output through pathways that have not been explicitly marked as acceptable. In the language of [28], ASVs serve as a generic measure of *unresolved discrimination*. To be clear, ASVs do not compute the direct causal effect of gender on admission, as say in [30], but instead measure the incremental effect of gender on the model's prediction, through paths not containing resolving variables.

Importantly, one need not know the causal graph underlying the data in order to operate in this framework. Instead, one simply needs to list the sensitive attributes (e.g. gender or race) and resolving

[6] Whether a process is deemed fair is context dependent. Depending on the situation, this could be a question of legal compliance, company policy, or a team's moral compass.

variables (e.g. department choice or auto-collision history) for a given problem, then compute the corresponding ASVs with w_{fair}. This makes the ASV approach to measuring causal fairness highly amenable to policy decisions articulated in simple language.

4.3 Data types with intrinsic ordering

Next we demonstrate that ASVs offer a natural model explanation, addressing the sequential incrementality of the prediction, when the data being modelled has an intrinsic ordering. This is the case, for example, for both time series and language processing. To show this, we perform experiments using the Epileptic Seizure Recognition data set [42] from the UCI ML repository [36]. The data contains about 12k electroencephalogram (EEG) signals, each consisting of 178 time steps (features), which span 1 second of measurement, and a label indicating whether or not seizure activity was recorded. See Fig. 4(a) for an example EEG signal from each class.

We trained a recurrent-neural-network (RNN) classifier on the EEG data; see App. B.3 for details. The data has an 80/20 class balance and the RNN classifier achieves 98% accuracy. This is the model we aim to understand using ASVs.

To set a baseline, we first computed global (symmetric) Shapley values for this model. We did this by sampling the sum in Eq. (2), using the value function of Eq. (3), and aggregating according to Eq. (5). This baseline model explanation is labelled "Off data manifold" in Fig. 4(b). This explanation attributes significant feature importance across all time steps of the signal. This result is expected from the perspective of Axiom 4 (Symmetry) satisfied by Shapley values: as each EEG signal is an arbitrary one second snapshot of continuous brain activity, no individual region of the signal (e.g. the beginning or end) should be much more predictive of an epileptic seizure than any other. In this sense, ordinary Shapley values "spread out" the model explanation over the predictive features.

ASVs allow us to take advantage of the natural ordering of the time series to obtain a sparser explanation of the model's behaviour. To compute ASVs we sampled the sum in Eq. (13), used the value function of Eq. (9), and aggregated according to Eq. (5). For $p(x'_{\bar{S}}|x_S)$, we trained an RNN data imputer as described in App. B.3. For $w(\pi)$, we placed nonzero weight only on the trivial permutation that places features in time order:

$$w_{\text{ordered}}(\pi) = \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ for all } i < j \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

ASVs computed with w_{ordered} thus indicate the additional predictivity gained by time step t assuming all steps $t' < t$ are already known. The corresponding values are labelled "ASV" in Fig. 4(b).

Thus, ASVs concentrate the feature importance into the beginning of the time series. In particular, the ASVs drop by a factor of roughly 100 after the first 25 time steps, while the Shapley values do not even change by a full order of magnitude over the entire series. This means that time steps 26 through 178 offer little additional predictive power once time steps 1 through 25 are already known (we demonstrate this assertion precisely in the next section). Since the same is true of any width-25 window in the time series, Shapley values are forced (by symmetry, Axiom 4) to spread the feature importance out along the entire sequence. This is the sense in which ASVs offer an explanation that is more sparse, and more fundamental to the sequential nature of the data, than the alternative.

4.4 Precise, verifiable feature selection

Finally, we demonstrate that ASVs support a direct interpretation in terms of the accuracy achievable by a model that uses only a subset of the data's features. This makes ASVs useful for feature-selection studies that aim to eliminate non-predictive features from a data set.

To explore this mathematically, suppose we define ASVs by weighting permutations with

$$w(\pi) = \frac{1}{|U|!\,|V|!} \times \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ for all } i \in U \text{ and all } j \in V \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

for some partition $U \sqcup V = N$ of all the data's features into disjoint subsets. Then the sum of global ASVs over features in U gives the accuracy achieved by the model using U if it is forced to marginalise

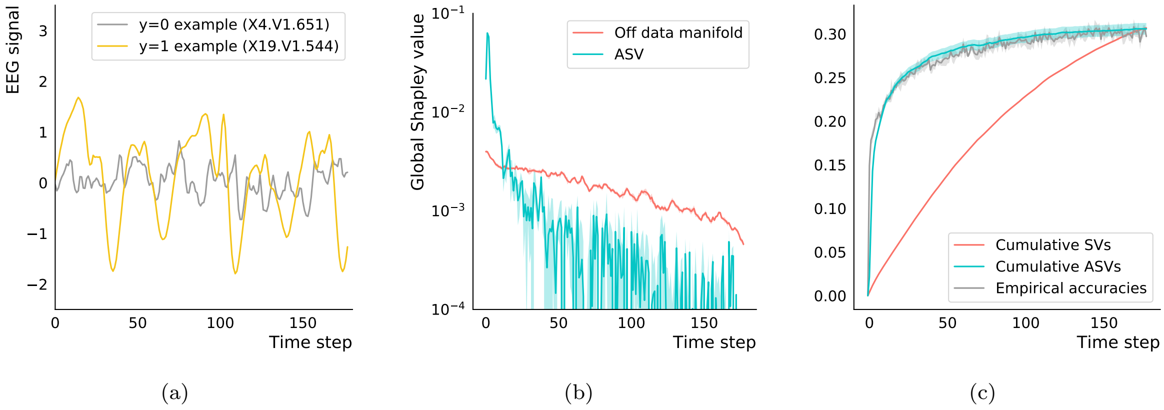


Figure 4: (a) Two example EEG signals from Epileptic Seizure Recognition data [42]. The gray example is benign, and the yellow signal corresponds to seizure activity. (b) Global explanations of a model trained on this data set. Shapley values, labelled "Off manifold", place all time steps on equal footing and thus distribute significant feature importance across the entire series. ASVs indicate the additional predictivity gained by time step t assuming all steps $t' < t$ are already known; these values concentrate attribution towards the beginning of the series and offer a sparser explanation of the model's predictions. (c) Cumulative values corresponding to the individual values in Fig. 4(b). At each time step t, these are compared to the empirical accuracy, defined according to Eq. (26), obtained for a model trained on features $t' \leq t$. This shows that each ASV has a direct interpretation in terms of the model accuracy attributable to its corresponding feature.

over V:

$$\sum_{i \in U} \phi_f^{(w)}(i) = \mathcal{A}_f(U) - \mathcal{A}_f(\{\}) \tag{26}$$

where we have used notation defined in Eq. (7) with the on-manifold value function in Eq. (9); refer to the surrounding discussion for further interpretation of these quantities. The analogous sum over features in V gives the improvement in accuracy the model can achieve using features in V instead of marginalising over them:

$$\sum_{i \in V} \phi_f^{(w)}(i) = \mathcal{A}_f(U \sqcup V) - \mathcal{A}_f(U) \tag{27}$$

Aside from imperfections in the model-to-be-explained f and the learnt distribution $p(x'|x_U)$, i.e. in the non-parametric limit, the accuracy of f after marginalisation over features in V should be equal to the accuracy achievable by another model trained solely on $\mathcal{X}_U$. Thus we can interpret the ASVs of features in U as contributions to the accuracy achievable by a model trained only on $\mathcal{X}_U$. The ASVs of features in V are then contributions to the marginal increase in accuracy achievable by a model trained on $\mathcal{X}_U \sqcup \mathcal{X}_V$. Thus, for a feature-selection study, Eq. (27) represents the decrease in accuracy one could expect upon dropping the features in V.

Results analogous to Eqs. (26) and (27) would follow if we instead partitioned the features into many disjoint subsets $U_1 \sqcup U_2 \sqcup U_3 \sqcup \cdots = N$ in Eq. (25). We can demonstrate this using the model, data, and global explanations of Sec. 4.3. In that case, $w_{\text{ordered}}(\pi)$ requires π to be fully ordered, so the ASV of feature t corresponds to the marginal increase in accuracy that the model achieves by accepting feature t as input on top of features $t' < t$:

$$\phi_f^{(w)}(t) = \mathcal{A}_f(\{t' \leq t\}) - \mathcal{A}_f(\{t' < t\}) \tag{28}$$

Cumulative ASVs thus obey:

$$\sum_{t' \leq t} \phi_f^{(w)}(t') = \mathcal{A}_f(\{t' \leq t\}) - \mathcal{A}_f(\{\}) \tag{29}$$

due to the telescoping nature of the sum. We can directly test this assertion by (i) training a new model $f^{(t)}$ for each time step t in the data and (ii) comparing its accuracy, in the sense of Eq. (29), to the corresponding cumulative ASV.

We thus trained many models on the Epileptic Seizure Recognition data [42]. A separate model $f^{(t)}$, for each time step t, was trained on all previous time steps $t' \leq t$. See App. B.3 for a description of these models. We plotted the accuracy, in the sense of Eq. (29), of each of these models in gray in Fig. 4(c). We compared these empirical accuracies to cumulative Shapley values and cumulative ASVs in Fig. 4(c). The cumulative values in Fig. 4(c) correspond to the sum of the individual values from Fig. 4(b).

The close relationship between empirical accuracies and asymmetric values in Fig. 4(c) demonstrates that ASVs indeed have a precise, verifiable interpretation in terms of the model accuracy attributable to each feature in the data. This makes them useful for feature-selection studies as they allow the practitioner to avoid re-training many models (here $f^{(t)}$) on various subsets of the data's features. Instead, to compute ASVs, one must simply train a single model (f) as well as a probabilistic data imputer $p(x'_{\bar{S}}|x_S)$ (App. A). Deviations between empirical accuracies and cumulative ASVs in Fig. 4(c) are due to imperfections in the models $f^{(t)}$ and f as well as the data imputer $p(x'|x_S)$.

5 Conclusions

In this work, we introduced Asymmetric Shapley values, a mathematically principled framework for model-agnostic explainability that generalises the standard Shapley values in order to incorporate causal information relevant to the underlying data on which the model operates. We showed how this framework can be employed across multiple applications in modern machine learning: incorporating the causal structure of data into model explanations, testing for fairness amidst subtleties like resolving variables, constructing sequential explanations in the context of time-series data, and selecting important features without model retraining. We expect that these applications are but a sample of the possible ways in which ASVs can improve model-agnostic explainability. Indeed, it is our hope that lowering the barrier to incorporating causality into AI explainability will lead to the development of better models and the deployment of more trustworthy AI systems throughout society.

Acknowledgements

This work was developed and experiments were run on the Faculty Platform for machine learning. The authors benefited from discussions with Christiane Ahlheim, Tom Begley, Julian Berman, Laurence Cowton, Markus Kenusch, Omar Sosa Rodriguez, Ron Smith, Megan Stanley, and Naoki Yoshihara. C.R. is grateful to Birkbeck College for its hospitality.

References

[1] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

[2] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *International Conference on Knowledge Discovery and Data Mining*, 2016.

[3] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 2017.

[4] L. Breiman. Random forests. *Machine learning*, 2001.

[5] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 2008.

[6] A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, 2018. [arXiv:1801.01489].

[7] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 2010.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you: Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*, 2016.

[9] L. S. Shapley. A value for n-person games. In *Contribution to the theory of games*. 1953.

[10] S. Lipovetsky and M. Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 2001.

[11] I. Kononenko et al. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 2010.

[12] E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 2014.

[13] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy*, 2016.

[14] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[15] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles, 2018. [arXiv:1802.03888].

[16] C. Frye, L. Cowton, M. Stanley, and I. Feige. *Forthcoming*, 2019.

[17] P. Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 2010.

[18] J. Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 2010.

[19] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search.* MIT Press, 2nd edition, 2000.

[20] J. Pearl. *Causality: models, reasoning and inference.* Springer, 2000.

[21] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 2019.

[22] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, 2013.

[23] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining*, 2015.

[24] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2015. `[arXiv:1507.05259]`.

[25] H. Edwards and A. Storkey. Censoring representations with an adversary, 2015. `[arXiv:1511.05897]`.

[26] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 2016.

[27] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, 2012.

[28] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, 2017.

[29] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 2017.

[30] S. Chiappa. Path-specific counterfactual fairness. In *AAAI Conference on Artificial Intelligence*, 2019.

[31] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, 1944.

[32] J. Chen, L. Song, M. Wainwright, and M. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, 2018.

[33] J. Aas, M. Jullum, and A. Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values, 2019. `[arXiv:1903.10464]`.

[34] R. J. Weber. *Probabilistic values for games.* Cambridge University Press, 1988.

[35] D. Monderer and D. Samet. Variations on the shapley value. *Handbook of game theory with economic applications*, 2002.

[36] D. Dua and C. Graff. UCI machine learning repository, 2017. `[http://archive.ics.uci.edu/ml]`.

[37] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

[38] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores, 2016. `[arXiv:1609.05807]`.

[39] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 2017.

[40] S. Chiappa and W. S. Isaac. A causal bayesian networks viewpoint on fairness. In *International Summer School on Privacy and Identity Management*, 2018.

[41] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 1975.

[42] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 2001.

[43] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[44] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

[45] C. Ma, S. Tschiatschek, K. Palla, J. M. Hernández-Lobato, S. Nowozin, and C. Zhang. EDDI: efficient dynamic discovery of high-value information with partial VAE. In *International Conference on Machine Learning*, 2019.

[46] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

A Probabilistic data imputer

The value function of Eq. (9) underlies the ASVs presented in the experiments of Sec. 4. This value function refers to a distribution $p(x'|x_S)$ that cannot be reliably estimated empirically on high-dimensional or continuous data. We refer to any model of $p(x'|x_S)$ as a probabilistic data imputer. A general method to learn $p(x'|x_S)$ using variational inference is presented in [16] and briefly summarised here for completeness.

Our probabilistic data imputer has 3 components: a stochastic encoder $q_\phi(z|x)$, a stochastic decoder $p_\theta(x|z)$, and a stochastic masked encoder $r_\psi(z|x_S)$. By themselves, the encoder q_ϕ and decoder p_θ comprise a variational autoencoder (VAE) [43, 44]. The VAE objective function is appended with terms that reward the masked encoder r_ψ for encoding the partial input x_S into a similar region of latent (z) space as the encoder q_ϕ maps the full input x. This is justified mathematically in [16]. We used neural networks to model q_ϕ, p_θ, and r_ψ with architectures reported in App. B.

The objective function for the data imputer includes a VAE term and a masked-autoencoder term:

$$\begin{aligned} \mathcal{L}_{\text{imputer}} &= \alpha\, \mathcal{L}_{\text{VAE}} + (1-\alpha)\, \mathcal{L}_{\text{MAE}} \qquad (30) \\ \mathcal{L}_{\text{VAE}} &= \mathbb{E}_{q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] - \mathcal{D}_{\text{KL}}\Big(q_\phi(z|x)\,||\,p(z)\Big) \\ \mathcal{L}_{\text{MAE}} &= \mathbb{E}_{q_\phi(z|x)}\Big[\log p_\theta(x_{\bar{S}}|z)\Big] - \mathcal{D}_{\text{KL}}\Big(q_\phi(z|x)\,||\,r_\psi(z|x_S)\Big) \end{aligned}$$

The prior distribution $p(z)$ is fixed to an independent unit normal distribution for each component of z, and we used $\alpha = 1/2$ in all experiments.

Note that $r_\psi(z|x_S)$ requires a variable-size input x_S. This was not a problem for the experiments of Secs. 4.3 and 4.4, in which x_S is a time series and $r_\psi(z|x_S)$ was modelled with an LSTM, which accommodates variable-size input. To model $r_\psi(z|x_S)$ in the experiments of Secs. 4.1 and 4.2, we used the "PNP" method from [45], which works as follows.

Let n denote the dimensionality of x and suppose we want to feed x_S into the masked encoder. First, we multiply each element x_i of x_S by an embedding vector e_i of dimension n. Next we feed that product through a common nonlinearity with n hidden units to obtain $h(x_i * e_i)$. Finally we aggregate the results: $g_S = \sum_{i \in S} h(x_i * e_i)$. This results in n-dimensional g_S regardless of the cardinality of S. This embedded masked input is then fed through several hidden layers as described in Apps. B.1 and B.2. The components of h and each e_i are considered part of r_ψ and learnt during optimisation of the objective Eq. (30).

B Details of experiments

B.1 Experiment on Census Income data

For the experiment of Sec. 4.1, we used the Census Income (or "Adult") data set from the UCI ML repository [36], ignoring the "fnlwgt" final weight feature (intended to make the sample more representative) in this analysis for simplicity. The model-to-explain f was a densely connected neural network, with 2 hidden layers of 100 units. Using a 75/25 train/test split, the model was trained with default sklearn settings and early stopping on a validation fraction of 25%. While the data has a 76/24 class balance, the model f achieves 84.7% test-set accuracy. All results shown in Sec. 4.1 were computed on the held-out test set $\mathcal{X}_{\text{test}}$.

Three variants of global Shapley values appear in Fig. 2. Each is an aggregation of local values defined according to Eq. (5). The first variant, labelled "Off data manifold", is the standard one, defined by the sum over coalitions in Eq. (2) and the value function in Eq. (3). To obtain a Monte Carlo estimate of each global off-manifold Shapley value, we independently sampled two data points $x, x' \in \mathcal{X}_{\text{test}}$ and a permutation $\pi \in \Pi(N)$, then plugged these quantities into Eqs. (2), (3), and (5). We repeated this with 10^6 samples, plotting the resulting mean as the bar length in Fig. 2 and the standard error of the mean as the error bar.

The "On data manifold" and "ASV" results of Fig. 2 are similar Monte Carlo estimates on 10^6 samples. Global on-manifold Shapley values are defined by the sum in Eq. (2), the value function in Eq. (9), and the aggregation procedure in Eq. (5). The quantity $p(x'|x_S)$ appears in this value function.

We modelled this distribution using a probabilistic data imputer, described generally in App. A. We used dense neural networks for the encoder, decoder, and masked encoder, each with 2 hidden layers of 100 units. The data imputer was trained using Adam [46] for optimisation, a batch size of 128, and early stopping with a validation fraction of 25% and patience of 20 epochs.

The global ASVs appearing in Fig. 2 are defined by the sum in Eq. (13), the value function in Eq. (9), and the aggregation procedure in Eq. (5). The data imputer just described was used to estimate these values as well.

B.2 Experiment on synthetic college admissions data

For the experiment of Sec. 4.2, we used two synthetic data sets, which we refer to as "fair" and "unfair", with data generating processes described qualitatively in Figs. 3(a) and 3(b). In both data sets, gender, X_1, is a binary random variable, with $X_1 = 0$ for women and $X_1 = 1$ for men. It is drawn according to

$$P(X_1 = 1) = 1/2 \tag{31}$$

Test score, X_2, is a normally distributed random variable:

$$x_2 \sim N(0, 1) \tag{32}$$

Department choice, X_3, is a binary variable drawn differently for women and men:

$$\begin{aligned} P(X_3 = 1|X_1 = 0) &= 0.8 \\ P(X_3 = 1|X_1 = 1) &= 0.2 \end{aligned} \tag{33}$$

so that women mostly apply to department $X_3 = 1$ and men to $X_3 = 0$. The outcome of admission, Y, is a binary variable drawn differently for the two data sets. In the fair case,

$$P(Y = 1|X_2 = x_2, X_3 = x_3) = \text{sigmoid}(x_2 + 2\,x_3 - 1) \tag{34}$$

so that department $X_3 = 1$ is more competitive than $X_3 = 0$. In the unfair case, admission is additionally based on (binary) unreported referrals, X_4, which are more prevalent for men than for women:

$$\begin{aligned} P(X_4 = 1|X_1 = 0) &= 1/3 \\ P(X_4 = 1|X_1 = 1) &= 2/3 \end{aligned} \tag{35}$$

While X_4 is not reported in the unfair data set, it has an important effect on admissions:

$$P(Y = 1|X_2 = x_2, X_3 = x_3, X_4 = x_4) = \text{sigmoid}(x_2 + 2\,x_3 + 2\,x_4 - 2) \tag{36}$$

Models-to-explain $f^{(\text{fair})}$ and $f^{(\text{unfair})}$ were fit to the two synthetic data sets. For each we used a densely connected network with 2 hidden layers of 10 units. Using a 75/25 train/test split, each model was trained with default sklearn settings and early stopping on a validation fraction of 25%. While each data set has a 50/50 class balance, $f^{(\text{fair})}$ and $f^{(\text{unfair})}$ achieve 73.6% and 73.2% test-set accuracy, respectively. All results in Sec. 4.2 were computed on the held-out test sets $\mathcal{X}_{\text{test}}^{(\text{fair})}$ and $\mathcal{X}_{\text{test}}^{(\text{unfair})}$.

A probabilistic data imputer (App. A) was also fit to each synthetic data set, $\mathcal{X}_{\text{train}}^{(\text{fair})}$ and $\mathcal{X}_{\text{train}}^{(\text{unfair})}$. We used dense neural networks for the encoder, decoder, and masked encoder, each with 2 hidden layers of 20 units. The data imputer was trained using Adam [46] for optimisation, a batch size of 128, and early stopping with a validation fraction of 25% and patience of 20 epochs.

The global ASVs of Fig. 3 are defined by the sum in Eq. (13), the value function in Eq. (9), and the aggregation procedure in Eq. (5). We obtained Monte Carlo estimates of each sum of integrals as described in App. B.1, using 10^6 samples from each of $\mathcal{X}_{\text{test}}^{(\text{fair})}$ and $\mathcal{X}_{\text{test}}^{(\text{unfair})}$. Bar lengths correspond to means and error bars to standard error of the means.

B.3 Experiments on Epileptic Seizure Recognition data

The experiments of Secs. 4.3 and 4.4 were performed on the Epileptic Seizure Recognition data set [42] from the UCI ML repository [36]. The model-to-explain f was a recurrent neural network: an LSTM with 20-dimensional hidden state. Using a 75/25 train/test split, the model was trained using Adam [46] for optimisation, a batch size of 128, and early stopping with a validation fraction of 25% and patience of 20 epochs. While the data has an 80/20 class balance, the model f achieved 98.3% test-set accuracy. All results shown in Sec. 4.3 were computed on the held-out test set $\mathcal{X}_{\text{test}}$.

The global "Off data manifold" (symmetric) Shapley values in Fig. 4 are defined by the sum in Eq. (2), the value function in Eq. (3), and the aggregation procedure in Eq. (5). We obtained Monte Carlo estimates of each sum of integrals as described in App. B.1, here using 10^5 samples from $\mathcal{X}_{\text{test}}$. Central values in Figs. 4(b) and 4(c) represent means, with shaded uncertainty bands for the standard error of the means.

The global ASVs in Fig. 4 are defined by the sum in Eq. (13), the value function in Eq. (9), and the aggregation procedure in Eq. (5). Monte Carlo estimates again used 10^5 samples. The distribution $p(x'|x_S)$ appearing in Eq. (9) was modelled with a probabilistic data imputer, as described in App. A. We used LSTMs for the encoder, decoder, and masked encoder, each with a 20-dimensional hidden state. The data imputer was trained using Adam [46] for optimisation, a batch size of 512, and early stopping with a validation fraction of 25% and patience of 100 epochs. We varied hyperparameters (for LSTM dimension, batch size, and patience) up and down by a factor of 2 without seeing significant changes in the results.

As described in Sec. 4.4, Fig. 4(c) displays the cumulative values corresponding the individual values in Fig. 4(b). Fig. 4(c) also displays a gray curve labelled "Empirical accuracies". For each point t on this curve, a model $f^{(t)}$ was trained to use the restricted time steps $t' \leq t$ to perform the binary classification task on $\mathcal{X}_{\text{train}}$. These models were defined and trained identically to the model-to-explain f described above. We computed $\mathcal{A}_{f^{(t)}}(\{t' \leq t\}) - \mathcal{A}_{f^{(t)}}(\{\})$ on $\mathcal{X}_{\text{test}}$; see Eq. (7) and the surrounding discussion. This is the quantity that appears in Fig. 4(c). It can be interpreted as the accuracy, above a randomised baseline, achieved by model $f^{(t)}$. We performed 5 trials for each value of t; means appear as the central curve and standard deviations as the shaded uncertainty band.