Multi-choice Explanations: A New Cooperative Game Structure for XAI

Daniel Fryer, David Lowing, Inga Strümke, Hien Duy Nguyen

Multi-choice Explanations: A New Cooperative Game Structure for XAI

Daniel Fryer*, David Lowing†, Inga Strümke‡ and Hien Nguyen*
*School of Mathematics and Physics, The University of Queensland, St Lucia. E-mail: {daniel.fryer, h.nguyen7}@uq.edu.au
†Industrial Engineering Research Department, CentraleSupélec, Université Paris-Saclay. E-mail: david.lowing@centralesupelec.fr
‡Department of Computer Science, The Norwegian University of Science and Technology. E-mail: inga.strumke@ntnu.no

***Abstract*—Cooperative game theorists propose the following attractive process: (1) capture the abstract *value* of each possible *coalition* of individuals, (2) write down some principles, or *axioms*, on how to distribute the value (e.g., allocate *importance* to features or parameters), and then, (3) find a set of allocations that satisfy the principles. The Shapley value has received much attention – but it is just one solution concept, satisfying one set of principles, in one class of games. It is popular among game theorists because the axioms, and the class of TU-games, are reasonable in game theory. In AI and ML, we should choose carefully what is reasonable for our own purposes. In this paper, we highlight solution concepts in the class of multi-choice games (MC-games). These are model agnostic, and unique to their own set of axioms, just like the Shapley value. This paper offers a general algorithm for constructing any MC-game framework with polynomial time complexity in the number of parameter levels, and an application of this algorithm that is transparent, and can be readily generalised to local explanation frameworks such as SHapley Additive exPlanations (SHAP).**

***Index Terms*—Shapley, feature importance, SHAP, polynomial regression, feature selection, game theory.**

I. Introduction

Interest in the Shapley value within explainable AI (XAI), underpins the popularity of axiomatic solutions and cooperative games, for gaining model agnostic insights into feature or parameter importances. However, the validity of such solutions depends on accurate interpretation of the resulting values and appropriate application of the game theoretic structure [7], [15], [18].

While our research is interconnected with explainable AI, we choose to tactfully avoid the complex task of defining what constitutes an explanation or interpretation *per se* (refer to [5], [19], [21]). Rather, our focus is on fostering comprehension and transparency. Our objective, here, is to introduce a comprehensible game theoretic framework.

We now semi-formally introduce multi-choice (MC) games as feature attribution frameworks, culminating in a generalized algorithm for their usage (Algorithm 1): In the general setting, we have a random sample $\boldsymbol{Z}^\top = (Z_1, \ldots, Z_k)$ of data from some joint distribution (i.e., $\boldsymbol{Z} \sim P$), and we wish to model this distribution (e.g., for the purpose of prediction or inference) using a function $f_{\boldsymbol{\theta}}(\boldsymbol{Z})$ (such as a supervised or unsupervised learning model) that can be optimised over the domain of the parameter $\boldsymbol{\theta}$. In the supervised case, we may view the data as having a response variable $Y = Z_1$, and $k-1$ features $X_1 = Z_2, \ldots, X_{k-1} = Z_k$.

There is commonly at least one natural way to define submodels of $f_{\boldsymbol{\theta}}$, by marginalising [cite] or removing [cite] features, or changing parameters (e.g., reducing the depth of a decision tree, the number of learners in an ensemble, or the number of hidden layers in a neural network). In general, we are typically able to define a natural hierarchy $\mathcal{F}$ of submodels of $f_{\boldsymbol{\theta}}$, from the simplest (i.e., null) submodel, to a maximum complexity model $f_{\boldsymbol{\theta}}$.

To represent submodels in $\mathcal{F}$, define a vector $\boldsymbol{s} = (s_1, \ldots, s_n) \in \{0,1\}^n$ whose ith element s_i indicates if an indexed parameter (or feature) i is included, $s_i = 1$, or excluded, $s_i = 0$, in a submodel $f_{\boldsymbol{s},\boldsymbol{\theta}} \in \mathcal{F}$. We refer to each $\boldsymbol{s}$ as a participation profile. See Figure 1a for an example of an ordering of binary participation profiles. If a real-valued function v is defined, such that $v(\boldsymbol{s})$ represents some evaluation of the model $f_{\boldsymbol{s},\boldsymbol{\theta}}(\boldsymbol{Z})$, then the pair $(\{0,1\}^n, v)$ can be interpreted as a transferable utility (TU) game. The Shapley value [1] of any TU-game summarises the entire hierarchy $v(\mathcal{F})$, in a vector $(\mathrm{Sh}_1, \ldots, \mathrm{Sh}_n)$, where Sh_i is taken to measure the overall influence of setting $s_i = 1$.

However, many submodel hierarchies would be better suited to non-binary profiles. For example, consider a neural network with n hidden layers, having s_ℓ units in the ℓth layer. We can define a non-binary profile $\boldsymbol{s} = (s_1, \ldots, s_n)$, where $\boldsymbol{s}$ only counts the number of units in each layer. Take the simple case of $n = 2$ layers, and 2 units in each layer. Then, the profile $\boldsymbol{s} = (2,1)$ represents a subnetwork with 2 units in layer 1, and 1 unit in layer 2. The value $v(2,1)$ does not distinguish between the two units in layer 1. Since the participation profiles are non-binary, this situation represents an MC-game, rather than a TU-game (see Table I).

TABLE I: The subnetworks of $\boldsymbol{s} = (2,2)$, represented as a multi-choice (MC) game. The value $v(\boldsymbol{s})$ of each subnetwork is assumed to be defined, and we wish to assign a worth φ_{ij} to the number of units j in each hidden layer i, for each $j > 0$.

	$s_2 = 0$	$s_2 = 1$	$s_2 = 2$	worth of s_1
$s_1 = 0$	$v(0,0)$	$v(0,1)$	$v(0,2)$	0
$s_1 = 1$	$v(1,0)$	$v(1,1)$	$v(1,2)$	φ_{11}
$s_1 = 2$	$v(2,0)$	$v(2,1)$	$v(2,2)$	φ_{12}
worth of s_2	0	φ_{21}	φ_{22}	

A submodel hierarchy with indices $\boldsymbol{s} \in \mathcal{M} \subset \mathbb{N}_0^n$, together with a function $v : \boldsymbol{s} \mapsto \mathbb{R}$, defines an MC-game $(\mathcal{M}, v)$. There must be a maximum index $\boldsymbol{m} \geq \boldsymbol{s}$, corresponding to the full (maximum complexity) model $f_{\boldsymbol{m},\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}$. Notice that a TU-game is the special case where $\mathcal{M} = \{0,1\}^n$. An MC-game solution concept is essentially a method to marginalise the matrix of profiles $\mathcal{M}$ (see the margins of Table I, cf. [15]). In Figures 1a to 1c, we compare the ordering of binary index vectors (via Hasse diagrams), to the corresponding non-binary index vectors.

More formally, an MC-game framework can be constructed using Algorithm 1. Here, we focus on global explanations, where there is a single function $v : \mathcal{F} \to \mathbb{R}$. However, for local explanations, such as SHAP [17], the algorithm generalises with a collection of k functions $v_i,\ i = 1, \ldots, k$ (e.g., where index i points to a distinct observation in a sample of size k).

Algorithm 1: Constructing an MC-game framework

Data: A sample $\boldsymbol{Z} \sim P$, for some probability distribution P.

1 Define a hierarchy of submodels $\mathcal{F} = \{f_{\boldsymbol{s},\boldsymbol{\theta}} : \boldsymbol{s} \in \mathcal{M}\}$, where $\mathcal{M} = \{\boldsymbol{s} \in \mathbb{N}_0^n : \mathbf{0} \leq \boldsymbol{s} \leq \boldsymbol{m}\}$, for some maximum index $\boldsymbol{m} \in \mathbb{N}^n$. Here, $\boldsymbol{s}$ controls the complexity of the submodels, via n model parameters (or features), and $\boldsymbol{m}$ represents the full model.

2 Define a real-valued characteristic function $\hat{v}(\boldsymbol{s} \mid \boldsymbol{Z})$, that estimates the worth $v(\boldsymbol{s})$ (e.g., goodness-of-fit, or predictive performance) of each target model $f_{\boldsymbol{s},\boldsymbol{\theta}}$. It is required that $v(\mathbf{0}) = 0$.

3 Choose an MC-game value φ, interpreting v as an MC-game, $(\mathcal{M}, v)$. The value φ may be chosen for its appeal in satisfying a set of reasonable interpretability axioms.

4 Return a payoff matrix $\boldsymbol{\varphi}(\boldsymbol{m}, \hat{v})$, whose element φ_{ij} is interpreted as the estimated influence of parameter i at complexity level j, on the measurement $v(\boldsymbol{m})$. The dimension of $\boldsymbol{\varphi}(\boldsymbol{m}, \hat{v})$ is $n \times m_\top$, where $m_\top$ is a maximal coordinate of $\boldsymbol{m}$.

Algorithm 1, being a global method, avoids a well-known (and debated [12]) issue: by permuting the input data, or computing marginal expectations, many local permutation methods, such as KernelSHAP, create semi-synthetic data that fall outside the generative joint distribution [6], [9], [14]. Note, Step 1 has a significant impact on the interpretation of the resulting value. That is, when defining the submodel hierarchy, one has the opportunity to map submodels to player-level pairs that act on v either additively, or independently, or both (see [15] for a detailed discussion of *relatively inessential* characteristic functions, in the context of TU-games).

II. The general setting of MC-games

We now formally present the setting of MC-games. Let $N = \{1, \ldots, n\}$ be a fixed player set. Each player $i \in N$ has a finite set of pairwise distinct participation levels $\{0, \ldots, m_i\}$. We use $m_\top = \max\{m_i : i \in N\}$ to denote a maximal participation level. An element $\boldsymbol{s} \in \mathcal{M} = \prod_{i \in N}\{0, \ldots, m_i\}$ is referred to as a participation profile. A player $i \in N$ participates at level $j \in \{0, \ldots, m_i\}$ in profile $\boldsymbol{s}$, if $s_i = j$. The set $\mathcal{M}$ endowed with the usual binary relation $\geq$ on $\mathbb{R}^n$ induces a (complete) lattice (see Figure 1c).

An MC-game on N is a couple $(\mathcal{M}, v)$, where

$$v : \mathcal{M} \to \mathbb{R}$$

is a function, with $v(\mathbf{0}) = 0$, describing the worth $v(\boldsymbol{s}) \in \mathbb{R}$ of each profile $\boldsymbol{s} \in \mathcal{M}$. Denote by $\mathcal{G}$ the class of MC-games. Clearly, TU-games form a subclass of $\mathcal{G}$.

A value $\boldsymbol{\varphi} : \mathcal{G} \to \mathbb{R}^{n \times m_\top}$ is a map that assigns a unique $(n \times m_\top)$-dimensional payoff matrix to each $(\mathcal{M}, v) \in \mathcal{G}$. Each $\varphi_{ij}(\mathcal{M}, v)$, $i \in N$, $j \leq m_\top$, represents the payoff obtained by player i for its j-th participation level. By convention, we fix $\varphi_{ij}(\mathcal{M}, v) = 0$ for any $j > m_i$.

A. Values for MC-games

We present three distinct values for MC-games, which were introduced by [4], [20] and [16], called the DP, PZ and LT values, respectively. Each of these values extend the Shapley value to the framework of MC-games.

We define these values using their expression in terms of Harsanyi dividends. Pick any $(\mathcal{M}, v) \in \mathcal{G}$. The (multi-choice) Harsanyi dividends associated with $(\mathcal{M}, v)$ are recursively defined as

$$\forall \boldsymbol{s} \in \mathcal{M}, \quad \Delta_v(\boldsymbol{s}) = v(\boldsymbol{s}) - \sum_{\substack{t \leq \boldsymbol{s} \\ t \neq \boldsymbol{s}}} \Delta_v(t), \tag{1}$$

$$\text{and} \quad \Delta_v(\mathbf{0}) = 0.$$

The DP value was axiomatically characterized by [13], using four axioms (see Section V). Under the DP_{ij} value [4], player i receives a share of the Harsanyi dividend from each coalition in which it participates at least at level j. Each share received by player i decreases proportionally to the sum of all participation levels in the profile.

Definition 1. *For each $(\mathcal{M}, v) \in \mathcal{G}$, the* DP *value is defined as, $\forall i \in N, j \leq m_i$,*

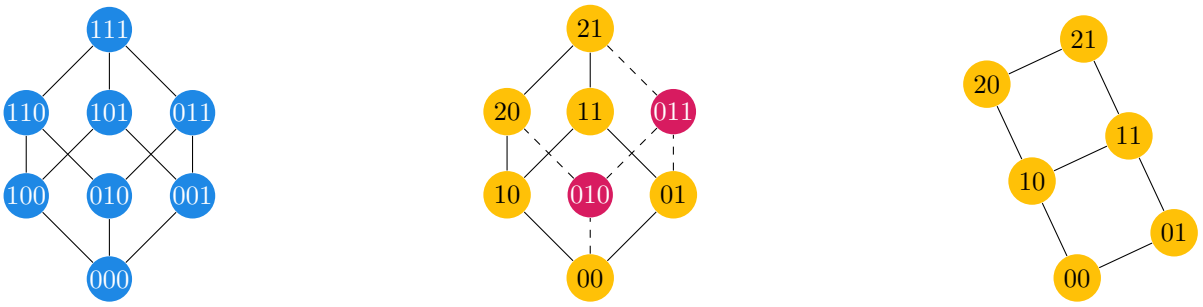
$$\text{DP}_{ij}(\mathcal{M}, v) = \sum_{\substack{\boldsymbol{s} \in \mathcal{M} \\ j \leq s_i}} \frac{\Delta_v(\boldsymbol{s})}{\|\boldsymbol{s}\|_1}. \tag{2}$$

The PZ value was axiomatically characterized by [20], using five axioms (see V). Under the PZ_{ij} value [20], player i receives a share of the Harsanyi dividend of each profile where player i participates at exactly level j. Each dividend is divided equally among all players that have non-zero participation.

Definition 2. *For each $(\mathcal{M}, v) \in \mathcal{G}$, the* PZ *value is defined as, $\forall i \in N, j \leq m_i$,*

$$\text{PZ}_{ij}(\mathcal{M}, v) = \sum_{\substack{\boldsymbol{s} \in \mathcal{M} \\ j = s_i}} \frac{\Delta_v(\boldsymbol{s})}{|\{k \in N : s_k > 0\}|}. \tag{3}$$

The LT value was axiomatically characterized by [16], using five axioms (see V). Under the LT_{ij} value [16], player i receives a share of the Harsanyi dividend of each profile for which player i, playing at level j, has the highest participation



(a) (b) (c)

Fig. 1: Hasse diagrams for the transition from (a) TU-game submodel ordering with $n = 3$, to (c) an MC-game submodel ordering with $\boldsymbol{m} = (2, 1)$. The transition is via (b) relabelling submodels (yellow) and removing redundant submodels (red).

level. Each dividend is divided equally among all players that participate at level j in the profile.

Definition 3. *For each $(\mathcal{M}, v) \in \mathcal{G}$, the* LT *value is defined as, $\forall i \in N, j \leq m_i$,*

$$\mathrm{LT}_{ij}(\mathcal{M}, v) = \sum_{\substack{\boldsymbol{s}\in\mathcal{M} \\ s_i=j \\ j=\max_{k\in N} s_k}} \frac{\Delta_v(\boldsymbol{s})}{|\{k \in N : s_k = j\}|}. \tag{4}$$

III. Polynomial regression feature explanations

In [10], the Shapley value is applied to multiple linear regression to decompose the coefficient of multiple correlation R^2 (a goodness-of-fit measure), for polynomials of degree 1. In this section, we trade the framework of [10] for an MC-game, by employing a map from participation levels to polynomials. We treat higher degree polynomial terms and interactions as greater levels of feature participation, rather than as separate features. This leads to a large reduction in computational complexity, and a decomposition that suits typical model explanation goals, as investigated further in Section IV-A.

A. Polynomial regression MC-game framework

Here, we introduce the MC-game polynomial regression framework, by parametrising a multiple polynomial regression submodel hierarchy and mapping it to an MC-game structure. Given a sample of *iid* observations from a random vector $\boldsymbol{Z}^\top = (Y, \boldsymbol{X}^\top)$ with $\boldsymbol{X}^\top = (X_1, \ldots, X_n)$, we define, for each $\boldsymbol{s} \leq \boldsymbol{m}$, $\boldsymbol{m} \in \mathbb{N}^n$, a multivariate polynomial $\mathcal{B}_{\boldsymbol{X}}(\boldsymbol{s})$, where the maximal degree of variable X_i is equal to s_i. That is,

$$\mathcal{B}_{\boldsymbol{X}}(\boldsymbol{s}) = \beta_0 + \sum_{i=1}^{n}\sum_{j=1}^{s_i} \beta_{ij} X_i^j + I(\boldsymbol{s}, d, p), \tag{5}$$

where $\beta_0, \beta_{ij} \in \mathbb{R}$ for each $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, s_i\}$, and $I(\boldsymbol{s}, d, p)$ consists of interaction terms whose complexity is controlled by $\boldsymbol{s}$, and $d, p \in \mathbb{N}$, as in,

$$I(\boldsymbol{s}, d, p) = \sum_{\boldsymbol{k}\in K(\boldsymbol{s},d,p)} \alpha_{\boldsymbol{k}} \prod_{i=1}^{n} X_i^{k_i}, \tag{6}$$

in which $\boldsymbol{k} = (k_1, \ldots, k_n)$ belongs to the set

$$\begin{aligned} K(\boldsymbol{s}, d, p) = \{\boldsymbol{k} \leq \boldsymbol{s} \,:\, & \sum_{i=1}^{n} \mathbb{1}(k_i > 0) \geq 2, \\ & \sum_{i=1}^{n} k_i \leq d,\ \max_{i\in\{1,\ldots,n\}}(k_i) \leq p\}. \end{aligned} \tag{7}$$

and $\alpha_{\boldsymbol{k}} \in \mathbb{R}$. The (fixed) hyperparameters d and p are employed in $I(\boldsymbol{s}, d, p)$ to limit the complexity of higher order interactions; we refer to these, respectively, as the degree and maximal power of the interactions. These hyperparameters are desirable in practical applications, where higher order interactions (and interactions between higher degree terms) may not be of interest. Note that all interaction terms vanish when $d < 2$; and the complete n-dimensional polynomial of degree d_0 is given by $d = m_1 = \ldots = m_n = d_0$, $p = d - 1$ (and occurs when $\boldsymbol{s} = \boldsymbol{m}$).

Example 1. *The parameters $d = m_1 = m_2 = 3$, $p = 2$ give the complete two-dimensional polynomial of degree 3,*

$$\begin{aligned} \mathcal{B}_{\boldsymbol{X}}(3,3) = \ & \beta_0 + \beta_{10}X_1 + \beta_{20}X_1^2 + \beta_{30}X_1^3 \\ & + \beta_{01}X_2 + \beta_{02}X_2^2 + \beta_{03}X_2^3 \\ & + \alpha_{11}X_1X_2 + \alpha_{21}X_1^2X_2 + \alpha_{12}X_1X_2^2. \end{aligned} \tag{8}$$

Example 2. *For $d, p = 3, 2$ and $\boldsymbol{m} \geq (2, 2, 2)$, the full interaction term component is,*

$$\begin{aligned} I(\boldsymbol{m}, 3, 2) = & \alpha_{110}X_1X_2 + \alpha_{101}X_1X_3 + \alpha_{011}X_2X_3 + \\ & \alpha_{210}X_1^2X_2 + \alpha_{102}X_1X_3^2 + \alpha_{021}X_2^2X_3 + \\ & \alpha_{120}X_1X_2^2 + \alpha_{201}X_1^2X_3 + \alpha_{012}X_2X_3^2 + \\ & \alpha_{111}X_1X_2X_3. \end{aligned} \tag{9}$$

To each polynomial $\mathcal{B}_{\boldsymbol{X}}(\boldsymbol{s})$, we may associate a linear (polynomial regression) model

$$Y = \mathcal{B}_{\boldsymbol{X}}(\boldsymbol{s}) + \varepsilon, \tag{10}$$

where ε has a distribution with zero mean and finite variance. The coefficients β_{ij} and $\alpha_{\boldsymbol{k}}$ can then be estimated via ordinary least squares. For each $\boldsymbol{s} \leq \boldsymbol{m}$, the goodness-of-fit (e.g.,

$R^2_{\boldsymbol{Z}}(\boldsymbol{s})$, the coefficient of multiple correlation of the sub-model (10) with sub-polynomial $\mathcal{B}_{\boldsymbol{X}}(\boldsymbol{s})$) for the corresponding $\mathcal{B}_{\boldsymbol{X}}(\boldsymbol{s})$ can be collected in a characteristic function

$$v_{\boldsymbol{Z}} : \prod_{i\in n}\{0,\ldots,m_i\} \to \mathbb{R}. \tag{11}$$

See the numerical study in Section IV-A.

B. Computational complexity of framework

For simplicity, let $\boldsymbol{m} = \boldsymbol{m}_0 = \{d_0\}^n$, where $d_0 \in \mathbb{N}$. The MC-game framework requires an evaluation of each submodel indexed by the domain of the characteristic function (11). The number of evaluations is therefore

$$|\mathrm{dom}(\nu_{\boldsymbol{Z}})| = \prod_{i=1}^{n}(m_i+1) = (d_0+1)^n. \tag{12}$$

In contrast, the TU-game framework of [10] requires evaluation of all polynomials constructed by removal of terms. Consider the best case (no interaction terms) and worst case (complete polynomial) scenarios. Let T_{min} denote the number of terms in the full polynomial without interaction terms (i.e., $d = 1$) and let T_{max} be the number of terms in the complete n-dimensional polynomial of degree d_0. Then,

$$\begin{aligned} T_{\text{min}} &= \sum_{i=1}^{n} m_i = nd_0, \\ T_{\text{max}} &= nd_0 + |K(\boldsymbol{m}_0, d_0, d_0-1)| \\ &= \binom{n+d_0}{d_0} - 1 \in [n^{d_0}, (n+d_0)^{d_0}] \end{aligned} \tag{13}$$

The number of submodel evaluations in the TU-game is between $2^{T_{\text{min}}}$ and $2^{T_{\text{max}}}$. It follows that the TU-game time complexity, even in the interaction-free case $2^{T}_{\text{min}} = 2^{nd_0}$, grows exponentially in the degree d_0, while the MC-game approach always grows polynomially in d_0, as in (12), regardless of interaction terms.

IV. Comparing game frameworks

A typical application for a feature or parameter attribution method is to find an interesting dataset, build a popular or useful model, compute some measure, attribute it to the features, and then interpret the attributions, making some argument that the attributions are reasonable or revealing (see [8], and cf. [2]).

A prevalent view in the cooperative game theory literature is that axiomatic characterisations, while abstract and appealing, are difficult to justify without a clear application context (for a detailed account, see [24]). In this view, the behaviour of a solution concept should be explored in a collection of relevant and transparent settings. With this in mind, the study in this section compares the DP, PZ and LT values to the Shapley value via simulation, across selected examples.

A. Experimental results

Here, via simulation, we highlight the differences between the TU-game and MC-game frameworks, in the context of polynomial regression. In each example, we use a sample size of 1000, draw each feature independently from a given uniform distribution, with a standard normal error term $\varepsilon \sim N(0,1)$, and compute values for each solution concept, with the coefficient of multiple correlation R^2 as characteristic function. For the MC-game, we use $\boldsymbol{m} = (3,3)$, $d = 2$ and $p = 1$, across all three models. For example, in this MC-game, the profiles in which player 1 participates at level 1 are $\boldsymbol{s} = (1,0),(1,1),(1,2),(1,3)$. These profiles correspond to the polynomials,

$$\begin{aligned} \mathcal{B}_{\boldsymbol{X}}(1,0) &= \beta_1 X_1, \\ \mathcal{B}_{\boldsymbol{X}}(1,1) &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2, \\ \mathcal{B}_{\boldsymbol{X}}(1,2) &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_2^2, \\ \mathcal{B}_{\boldsymbol{X}}(1,3) &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_2^2 + \beta_5 X_2^3. \end{aligned}$$

For the TU-game, in each case, we treat each term of the polynomial $\mathcal{B}_{\boldsymbol{X}}(3,3)$ as a separate player – the corresponding terms are shown in Table III.

We repeat each simulation 100 times, to achieve a mean and 95% quantile interval (QI) for the mean, for each estimated value. Chosen to highlight key differences in the two classes of games, the three simple models that we investigate are:

$$Y = X_1 + X_2 + \varepsilon, \qquad X_i \sim \mathrm{U}(-5,1), \tag{M1}$$
$$Y = X_1 + X_2 + X_1X_2 + \varepsilon, \qquad X_i \sim \mathrm{U}(4,5), \tag{M2}$$
$$Y = X_1 + X_2 + X_2^2 + \varepsilon, \qquad X_i \sim \mathrm{U}(-5,1). \tag{M3}$$

TABLE II: MC-game values estimated via simulation, for the three models (M1) to (M3). The 95% QI widths were all below 0.06. Greyed cells correspond to profiles that introduce terms that do not appear in the generative model.

M	S	$\hat{\varphi}_{11}$	$\hat{\varphi}_{12}$	$\hat{\varphi}_{13}$	$\hat{\varphi}_{21}$	$\hat{\varphi}_{22}$	$\hat{\varphi}_{23}$
M1	DP	0.43	0	0	0.43	0	0
	PZ	0.43	0	0	0.43	0	0
	LT	0.43	0	0	0.43	0	0
M2	DP	0.42	0	0	0.42	0	0
	PZ	0.42	0	0	0.42	0	0
	LT	0.42	0	0	0.42	0	0
M3	DP	0.08	0	0	0.80	0.09	0
	PZ	0.08	0	0	0.71	0.19	0
	LT	0.08	0	0	0.71	0.19	0

TABLE III: TU-game Shapley values estimated via simulation, for the three models (M1) to (M3). The 95% QI widths were all below 0.02. Greyed cells correspond to terms that do not appear in the generative model.

M	X_1	X_1^2	X_1^3	X_2	X_2^2	X_2^3	X_1X_2
M1	0.15	0.11	0.09	0.15	0.11	0.09	0.15
M2	0.1	0.1	0.1	0.1	0.1	0.1	0.22
M3	0.04	0.03	0.02	0.22	0.31	0.3	0.06

The interaction terms are treated as separate players in the TU-game for Table III. Alternatively, it is possible to construct a TU-game framework in which, e.g., inclusion of the terms X_1, X_2 implies inclusion of their interaction X_1X_2. Such a construction necessitates more decisions, e.g., does inclusion of X_1 imply inclusion of X_1^2? and, do X_1^2, X_2^2 together imply the presence of X_1X_2? Designing such a framework for TU-games is outside of our scope. However, for transparency, we provide in Table IV the Shapley values computed in each case for a TU-game with no interaction term player. The corresponding MC-game values (with interaction degree $p = 1$) are equal, within two significant figures, to those in Table II.

TABLE IV: TU-game Shapley values estimated via simulation, for the three models (M1) to (M3), without including in the TU-game any interaction term. Greyed cells correspond to profiles that are not present in the generative model.

M	X_1	X_1^2	X_1^3	X_2	X_2^2	X_2^3
M1	0.18	0.14	0.11	0.18	0.14	0.11
M2	0.14	0.14	0.14	0.14	0.14	0.14
M3	0.03	0.03	0.02	0.24	0.33	0.32

B. Discussion of results

Typically, higher degree terms are added to a polynomial model only in the presence of all lower degree terms. The MC-game regression framework obeys this feature-degree model hierarchy by design. In this hierarchy, the greyed cells in Tables II to IV should all contain 0. The TU-games incorporate no such assumptions. For this reason, terms like X_1^2, X_1^3 and X_2^3 can enter the TU-game alone, and receive credit for their performance in the absence of lower level terms.

Regarding interaction terms, comparing (M1) to (M2) in Tables II and III, we see that, while the MC-game is uninformative about the interaction term, the Shapley value of the interaction term is maximal for both (M1) (which has no interaction term), and (M2). Again, this reflects the freedom for the interaction term to enter a profile independently of the other terms.

Between the three MC-game solution concepts (DP, PZ and LT), the choice of solution has played a minor role compared to the difference in results across the two game structures (TU and MC games). This highlights the significance of the game structure itself, though further work could determine appealing axioms for feature importance. Recent examples of such efforts include [26] and [11].

The DP, PZ and LT values are all equal to two significant figures, for (M1), (M2) in Table II. For (M3), the DP value concentrates more worth in the lower level φ_{21}. The conceptual differences are discussed further in Section V, where we compare the solution concepts axiomatically.

V. Axiomatic characterisations

The analytic and numerical approaches act as a sanity check for axiomatic approach – in this view, the axioms should inform us to search for applied counterexamples, while the examples should inform us to search for improved axioms.

Here, we ground the MC-game values axiomatically, with characterizations of the DP, PZ and LT values. Several axioms are discussed, some of them being straightforward extensions of axioms from TU-games to MC-games.

MC-Efficiency For each $(\mathcal{M}, v) \in \mathcal{G}$,

$$\sum_{i \in N} \sum_{j \leq m_i} \varphi_{ij}(\mathcal{M}, v) = v(\boldsymbol{m}).$$

MC-Additivity For each $(\mathcal{M}, v), (\mathcal{M}, w) \in \mathcal{G}$,

$$\boldsymbol{\varphi}(\mathcal{M}, v + w) = \boldsymbol{\varphi}(\mathcal{M}, v) + \boldsymbol{\varphi}(\mathcal{M}, w).$$

The next axiom requires that if the maximal participation level of each player reduces to a certain level j, then the payoff of each player for their j-th participation level should remain unchanged.

Independence of higher levels For each $(\mathcal{M}, v) \in \mathcal{G}$,

$$\forall j \leq m_\top, \quad \varphi_{ij}(\mathcal{M}, v) = \varphi_{ij}((\min\{j, m_k\})_{k \in N}, v).$$

If a player's participation at a certain level produces nothing, it seems reasonable to penalize them accordingly. [20] introduce an axiom indicating that non-productive participation levels should not receive anything from the value.

Non-productive level For each $(\mathcal{M}, v) \in \mathcal{G}$, if there is a $i \in N$ and $j \leq m_i$ that verifies

$$\forall \boldsymbol{s} \in \mathcal{M}, \quad v(\boldsymbol{s}_{-i}, j-1) = v(\boldsymbol{s}_{-i}, j),$$

then $\varphi_{ij}(\mathcal{M}, v) = 0$.

A player's participation level is said to be inessential if that player stops being productive past that level. [13] introduce an axiom indicating that inessential participation levels should not receive anything from the value.[1] The following is a weaker axiom than Non-productive level.

Inessential level For each $(\mathcal{M}, v) \in \mathcal{G}$, if there is a $i \in N$ and a $j \leq m_i$ that verifies

$$\forall \boldsymbol{s} \in \mathcal{M}, \forall l \geq j, \quad v(\boldsymbol{s}_{-i}, l-1) = v(\boldsymbol{s}_{-i}, l),$$

then $\varphi_{ij}(\mathcal{M}, v) = 0$.

A necessary level represents the level of participation of a player under which the worth of any profile is null. [25] introduce an axiom stating that two necessary levels should receive the same payoffs.[2]

Necessary level For each $(\mathcal{M}, v) \in \mathcal{G}$, if there are two $i, i' \in N$ and two $j \leq m_i$, $j' \leq m_{i'}$, verifying

$$\forall \boldsymbol{s} \in \mathcal{M}, s_i < j \text{ and/or } \boldsymbol{s}_{i'} < j', \quad v(\boldsymbol{s}) = 0,$$

then $\varphi_{ij}(\mathcal{M}, v) = \varphi_{i'j'}(\mathcal{M}, v)$.

[1] This axiom and the Non-productive level level axiom are both equivalent to the Null player axiom [23] when $\boldsymbol{m} = \mathbf{1}$.

[2] This axiom is equivalent to the Necessary player axiom [3] when $\boldsymbol{m} = \mathbf{1}$.

If a player has the same performances at two distinct participation levels, then it seems reasonable that these two levels obtain the same payoff.

Intra symmetry For each $(\mathcal{M}, v) \in \mathcal{G}$, if there is a $i \in N$ and $j, j' \leq m_i$ verifying, $\forall \boldsymbol{s} \in \mathcal{M}$,

$$v(\boldsymbol{s}_{-i}, j) - v(\boldsymbol{s}_{-i}, j-1) = v(\boldsymbol{s}_{-i}, j') - v(\boldsymbol{s}_{-i}, j'-1),$$

then $\varphi_{ij}(\mathcal{M}, v) = \varphi_{ij'}(\mathcal{M}, v)$.

[20] propose a straightforward extension of Anonymity from TU-games to MC-games. Denote by $\overline{\mathcal{G}}$ the sub-class of MC-games in which all player have the same number of participation levels. This is just for convenience, as fictive participation levels could be introduced to equalize the number of levels for each player.[3]

MC-Anonymity For each $(\mathcal{M}, v) \in \overline{\mathcal{G}}$, each $t \in \mathcal{M}$ and each order $\pi \in P(N)$, we define πt as $\pi t_{\pi(i)} = t_i$ for each $i \in N$, and πv as $\pi v(\pi t) = v(t)$. Then, it holds that

$$\varphi_{ij}(\mathcal{M}, v) = \varphi_{\pi(i)j}(m, \pi v).$$

[16] propose an axiom which guarantees that two players with the same performance at a given participation level should receive the same payoff for that level.[4]

MC-Symmetry For each $(\mathcal{M}, v) \in \mathcal{G}$, if there are two $i, i' \in N$ and a $j \leq m_i$, $j \leq m_{i'}$, verifying, $\forall \boldsymbol{s} \in \mathcal{M}$,

$$v(\boldsymbol{s}_{-i}, j) - v(\boldsymbol{s}_{-i}, j-1) = v(\boldsymbol{s}_{-i'}, j) - v(\boldsymbol{s}_{-i'}, j-1),$$

then $\varphi_{ij}(\mathcal{M}, v) = \varphi_{i'j}(\mathcal{M}, v)$.

[13] provide a characterization of the DP value.

Theorem 1 ([13]). *A value $\boldsymbol{\varphi}$ on $\mathcal{G}$ satisfies MC-Efficiency, MC-Additivity, Necessary level and Inessential level if and only if $\boldsymbol{\varphi} =$* DP.

[20] provide an axiomatic characterization of the PZ value.

Theorem 2 ([20]). *A value $\boldsymbol{\varphi}$ on $\overline{\mathcal{G}}$ satisfies MC-Efficiency, MC-Additivity, MC-Anonymity, Non-productive level and Intra symmetry if and only if $\boldsymbol{\varphi} =$* PZ.

We provide a characterization of the LT value.

Theorem 3. *A value $\boldsymbol{\varphi}$ on $\mathcal{G}$ satisfies MC-Efficiency, MC-Additivity, MC-Symmetry, Independence of higher levels and Inessential level if and only if $\boldsymbol{\varphi} =$* LT. *The proof is direct from Theorem 1 and Corollary 1 in [16].*

REFERENCES

[1] E. Algaba, V. Fragnelli, and J. Sánchez-Soriano, *Handbook of the Shapley value*. CRC Press, 2019.

[3]To permute the labels of two players it is necessary that these players have the same number of participation levels. Therefore, this axiom only holds on the class of MC-games in which all the players have the same maximal activity level. Clearly, this axiom is equivalent to the Anonymity axiom [22] when $\boldsymbol{m} = \mathbf{1}$.

[4]This axiom is equivalent to the Symmetry axiom [23] when $\boldsymbol{m} = \mathbf{1}$.

[2] E. Amparore, A. Perotti, and P. Bajardi, "To trust or not to trust an explanation: using leaf to evaluate local linear xai methods," *PeerJ Computer Science*, vol. 7, p. e479, 2021.

[3] S. Béal and F. Navarro, "Necessary versus equal players in axiomatic studies," *Operations Research Letters*, vol. 48, no. 3, pp. 385–391, 2020.

[4] J. Derks and H. Peters, "A Shapley value for games with restricted coalitions," *International Journal of Game Theory*, vol. 21, no. 4, pp. 351–360, 1993.

[5] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[6] C. Frye, D. de Mijolla, T. Begley, L. Cowton, M. Stanley, and I. Feige, "Shapley explainability on the data manifold," in *International Conference on Learning Representations*.

[7] D. Fryer, I. Strümke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," *IEEE Access*, vol. 9, pp. 144 352–144 360, 2021.

[8] T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen, "Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models," *Advances in neural information processing systems*, vol. 33, pp. 4778–4789, 2020.

[9] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance," *Statistics and Computing*, vol. 31, pp. 1–16, 2021.

[10] F. Huettner and M. Sunder, "Axiomatic arguments for decomposing goodness of fit according to shapley and owen values," *Electronic Journal of Statistics*, vol. 6, pp. 1239–1250, 2012.

[11] J. Janssen, V. Guan, and E. Robeva, "Ultra-marginal feature importance: Learning from data with causal guarantees," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 10 782–10 814.

[12] D. Janzing, L. Minorics, and P. Blöbaum, "Feature relevance quantification in explainable ai: A causal problem," in *International Conference on artificial intelligence and statistics*. PMLR, 2020, pp. 2907–2916.

[13] F. Klijn, M. Slikker, and J. Zarzuelo, "Characterizations of a multi-choice value," *International Journal of Game Theory*, vol. 28, no. 4, pp. 521–532, 1999.

[14] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with shapley-value-based explanations as feature importance measures," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5491–5500.

[15] I. Kumar, C. Scheidegger, S. Venkatasubramanian, and S. Friedler, "Shapley residuals: Quantifying the limits of the shapley value for explanations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 598–26 608, 2021.

[16] D. Lowing and K. Techer, "Marginalism, egalitarianism and efficiency in multi-choice games," *Social Choice and Welfare*, pp. 1–47, 2022.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[18] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models using shapley values," in *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*. Springer, 2020, pp. 17–38.

[19] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[20] H. Peters and H. Zank, "The egalitarian solution for multichoice games," *International Journal of Game Theory*, vol. 137, no. 1, pp. 399–409, 2005.

[21] A. D. Selbst and S. Barocas, "The intuitive appeal of explainable machines," *Fordham L. Rev.*, vol. 87, p. 1085, 2018.

[22] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[23] M. Shubik, "Incentives, decentralized control, the assignment of joint costs and internal pricing," *Management Science*, vol. 8, no. 3, pp. 325–343, 1962.

[24] W. Thomson, "On the axiomatic method and its recent applications to game theory and resource allocation," *Social Choice and Welfare*, vol. 18, no. 2, pp. 327–386, 2001.

[25] C. G. A. van den Nouweland, "Games and graphs in economic situations. tilburg university," 1993.

[26] I. Verdinelli and L. Wasserman, "Feature importance: A closer look at Shapley values and LOCO," *arXiv preprint arXiv:2303.05981*, 2023.