Improving the Perception of Fairness in Shapley-Based Allocations

Meir Nizri, Amos Azaria, Noam Hazon

Abstract

The Shapley value is one of the most important normative division schemes in cooperative game theory, satisfying basic axioms. However, some allocation according to the Shapley value may seem unfair to humans. In this paper, we develop an automatic method that generates intuitive explanations for a Shapley-based payoff allocation, which utilizes the basic axioms. Given any coalitional game, our method decomposes it to sub-games, for which it is easy to generate verbal explanations, and shows that the given game is composed of the sub-games. Since the payoff allocation for each sub-game is perceived as fair, the Shapley-based payoff allocation for the given game should seem fair as well. We run an experiment with 630 human participants and show that when applying our method, humans perceive the Shapley-based payoff allocation as more fair than the Shapley-based payoff allocation without any explanation or with explanations generated by other methods.

1 Introduction

An important research question in cooperative game theory is that of fair division: if agents form a coalition to achieve a common goal, how should they split the revenue or costs fairly? Various notions of fairness have been proposed in the cooperative game theory literature, like the Nash-Harsanyi bargaining solution [11, 12] or the nucleolus [23], but the Shapley value [24] has been termed the most important normative division scheme in cooperative game theory [26]. The Shapley value is based on the idea that the payoff of the game should be divided such that each agent's share is proportional to its contribution to the payoff. Specifically, the Shapley value is the average expected marginal contribution of one agent to all possible subsets of agents. Indeed, the Shapley value is considered fair since it is the only payoff allocation that satisfies the following four desirable axioms: efficiency, symmetry, null player property and additivity [13] (see Section 3 for formal definitions). These axioms admit strong normative and positive interpretations [4]. We note that there are several equivalent sets of axioms that characterize the Shapley value [18].

While the axioms satisfied by the Shapley value seem necessary, humans presented with an allocation according to the Shapley value may sometimes not observe it as fair (we experimentally support this claim in Section 7) [5, 6]. For example, consider the following game with three agents: r, l_1, and l_2, which is also known as the classical "glove game". Agents l_1 and l_2 have a left-glove and agent r has a right-glove. A pair of left and right gloves is worth \$12, but a single glove is worth nothing. If all agents collaborate, the Shapley value allocates \$8 to agent r and only \$2 to l_1 and \$2 to l_2. While it seems plausible that agent r should receive a higher payoff, a right-glove alone is worth nothing and thus, it may seem unfair that the payoff for this agent is 4-times more than each of the other agents. However, any other allocation would violate at least one of the axioms. It is thus desirable to increase human acceptance of the allocation according to the Shapley value, which can be achieved by providing explanations. In this paper, we develop an automatic method that generates intuitive explanations for a Shapley-based payoff allocation.

There are many possible ways for generating explanations for a Shapley-based payoff allocation. Indeed, Procaccia claimed that "the central role of axioms should be to help explain the mechanism's outcomes to participants" [22], and this direction has been successfully applied in the field of fair division by the *Spliddit* website [8]. We follow this idea, and build our explanations on top of the four axioms of the Shapley value.

Now, the essence of our explanation is that any game is decomposed into several sub-games that their Shapley allocation is easier to perceive as fair. Specifically, any sub-games is built such that all the agents are either null players or equivalent to one another, and the values are either all non-negative or all non-positive. According to the null player axiom each agent who is a null player should receive a payoff of 0, and according to the symmetry and efficiency axioms all other agents should equally share the total outcome, and thus the Shapley allocation in each sub-game is intuitively fair. For example, the "glove game" can be decomposed into few sub-games; in one of the sub-games, agent r obtains a value of $\$12$ when collaborating with l_1, but not when collaborating with l_2. When all three agents collaborate, they obtain a value of $\$12$. In this sub-game l_2 is a null player, and agents r and l_1 are equivalent. Thus, the Shapley allocation of $\$6$ to agent r, $\$6$ to agent l_1 and $\$0$ to agent l_2 is intuitively fair. Finally, following the additivity axiom, since the Shapley allocation of every sub-game is intuitively fair, and the sum of the Shapley allocations in each sub-game is equal to the Shapley allocation in the original game, then the latter is easier to perceive as fair. We note that this process follows the arguments in the proof of the uniqueness of the Shapley value [24].

Practically, we do not directly present the axioms to the users. Instead, our algorithm, which we termed *X-SHAP*, decomposes any coalitional game into several sub-games, and automatically generates a brief verbal explanation that accompanies each sub-game. For example, recall the sub-game of the "glove game" that we have previously mentioned. X-SHAP presents the sub-game to the user, and generates the following verbal explanation:

> *"In this scenario, l_2 does not contribute anything. r and l_1 are identical and always contribute the same. Therefore, the total revenue, which is $\$12$, should be equally divided between r and l_1, and thus, the fair division is $r : \$6, l_1 : \$6, l_2 : \$0$."*

Similarly, X-SHAP presents the other sub-games along with their explanations. X-SHAP finalizes its explanation by stressing out that since the sum of all the sub-games is the original game, the proposed division is fair as it is the sum of all the sub-games divisions.

In order to evaluate the performance of X-SHAP, we conducted a survey with human participants. The survey examined six coalitional games, representing a variety of scenarios. Each of the coalitional games was presented to the participants along with its Shapley payoff allocation as a suggestion for dividing the payoff among the agents. Then, each scenario was accompanied by one of the following: the complete explanations of X-SHAP, the decomposition into sub-games of X-SHAP without their verbal explanations, a heuristic decomposition into sub-games, a heuristic verbal explanation, a fixed general explanation of the benefits of the Shapley value, and no explanation at all. The participants were asked to rate the proposed allocation by indicating to what extent they agree or disagree that it is fair. Overall, 630 different people participated in the survey, each answering two different coalitional games with different explanation types. The explanations that were generated by X-SHAP achieved higher fairness ratings compared to the other explanations in all the games examined. This indicates that humans perceive the Shapley payoff allocation fairer if they receive X-SHAP's explanations.

To summarize, the main contribution of this paper is that it provides the first successful automatic method that generates customized explanations of the Shapley allocation for any coalitional game.

2 Related Works

Our work is related to the field of Explainable AI (XAI) [3, 9]. In a typical XAI setting, the goal is to explain the output of an AI system to a human. This explanation is important for allowing the human to trust the system, better understand, and to allow transparency of the system's output [1]. Other XAI systems are designed to provide explanations, comprehensible by humans, for legal or ethical reasons [7]. For example, an AI system for the medical domain might be required to explain

its choice for recommending the prescription of a specific drug [14]. Indeed, most of the work on XAI concerned the explanation of a machine learning based model. In this paper, we develop a system for explaining a solution concept that is based on a set of axioms. Our work can be also seen as an instance of x-MASE [16], explainable decisions in multiagent environments.

The work that is closest to ours, albeit in the context of voting, is by Cailloux and Endriss [2]. They propose a logic-based system for providing justifications for the outcome of a voting rule. They also develop an algorithm that automatically derives a justification for any outcome of the Borda rule. The algorithm's main idea is to decompose the preference profile into a sequence of sub-profiles, and use one of six axioms for providing explanations for the sub-profiles and for their combinations. This approach is further extended by Peters et al. [21], which investigate the required length of the sequence of explanations. Our approach for explaining the Shapley allocation is also based on axioms, and we also decompose the given coalitional game into a set of sub-games, which together compose an explanation for the given coalitional game.

Another work that analyzes a decomposition of a coalitional game in relation with the Shapley value is the paper by Stern and Tettenhorst [25]. They provide a new characterization of the Shapley value, by showing that a coalitional game can be decomposed into sub-games, one sub-game for each agent. They prove that the Shapley value equals the value of the grand coalition in each agent's sub-game. Similarly, de Clippel [4] provides a new axiomatization for the Shapley value by replacing the additivity axiom with the difference formula (DF) axiom. The DF axiom requires that each agent's payoff can be obtained by subtracting two functions: one function depending on the values of all sets that the agent belongs to, and the other depending on those that she does not belong to.

Spliddit [8] is a website implementing algorithms for various division tasks (e.g., rent division), which also explains how the outcomes satisfy certain fairness requisites. While the website enables users to compute the Shapley value in a ride-sharing context, it provides only a general explanation that states the benefits of the Shapley value. Our work can thus serve as an extension for Spliddit by providing customized explanations for the Shapley value.

The Shapley value can also be applied for increasing interpretability of a machine learning model. A common approach is SHAP [17]. For each prediction of any machine learning model, SHAP can calculate a list of values expressing the contribution of each feature the model considers to the prediction. It does so by simulating the behavior of the model to a coalitional game, where the features are "players" in the game and the predictions are the payoffs. In this setting, the problem becomes calculating the contribution of each "player", which can be done by calculating the average of the marginal contributions across all permutations for each feature.

3 Definitions

A coalitional game is defined by a pair (N, v), where N is a finite set of n agents and v is a function that associates every subset of N, a coalition, with a real value that represents the collective payoff its members can gain should they cooperate, i.e., $v : 2^N \rightarrow \mathbb{R}$. The function v is called the *characteristic function*. We assume that v always satisfies $v(\emptyset) = 0$. A characteristic function v is *super-additive* if for any pair of disjoint subsets S, T it holds that $v(S \cup T) \geq v(S) + v(T)$, and it is *sub-additive* if $v(S \cup T) \leq v(S) + v(T)$.

The main assumption in cooperative game theory is that the grand coalition N, which consists of all the agents, will form. A typical goal is then to allocate the value $v(N)$ among the agents in some fair way. A solution concept is a vector $\phi \in \mathbb{R}^N$ that represents the allocation to each agent $i \in N$.

The Shapley value is a solution concept that assigns a payoff to each agent according to their marginal contribution [24]. Formally, for each agent i,

$$Sh_i(N, v) = \sum_{S \subseteq N | i \in S} \frac{(|S| - 1)!(n - |S|)!}{n!}(v(S) - v(S \setminus \{i\})).$$

Shapley Value Axioms

The Shapley value is the only solution concept that simultaneously satisfies the following axioms [13].

Definition 1 (efficiency)**.** *The sum of all agents' payoff equals the grand coalition's value. That is,* $\sum_{i \in N} \phi_i(N, v) = v(N)$.

Definition 2 (symmetry)**.** *Two agents* i *and* j *are said to be equivalent if for any coalition* $S \subseteq N$ *that contains neither* i *nor* j*, it holds that* $v(S \cup \{i\}) = v(S \cup \{j\})$*. The symmetry axiom requires that equivalent agents receive the same payoff, i.e.,* $\phi_i(N, v) = \phi_j(N, v)$.

Definition 3 (null player)**.** *Agent* i *is said to be a null player if for every coalition* $S \subseteq N \setminus \{i\}$*, it holds that* $v(S \cup \{i\}) = v(S)$*. The null player axiom requires that the payoff for the null player will be* 0*, i.e.,* $\phi_i(N, v) = 0$.

Definition 4 (additivity)**.** *Given two coalitional games* (N, v) *and* (N, w)*, let* $v + w$ *be a function,* $v + w : 2^N \rightarrow \mathbb{R}$*, such that for every* $S \subseteq N$*,* $(v + w)(S) = v(S) + w(S)$*. The additivity axiom requires that the allocation to every agent* $i \in N$ *in the coalitional game* $(N, v + w)$ *satisfies* $\phi_i(N, v + w) = \phi_i(N, v) + \phi_i(N, w)$.

4 Coalitional Games that are Easy to Explain

While automatically generating explanations for any coalitional game may seem as a complex task, there exist coalitional games that it is possible to automatically generate compelling explanations for them. In this subsection we define easy-to-explain (ETX) games and show how to generate the appropriate explanations for them. In the next subsection, we use ETX games for generating explanations for *any* coalitional game.

Definition 5 (clean)**.** *A coalitional game* (N, v) *is said to be clean, if* v *is super-additive and consists of non-negative values only, or if* v *is sub-additive and consists of only non-positive values.*

Intuitively, a clean game represents a "common" scenario. Namely, a clean game can be associated with either a monetary revenue scenario or a taxation scenario. If a coalitional game consists of non-negative values only, then each coalition in this game may represent the collective revenue its members gain should they cooperate. It is common to assume that in a revenue scenario a collaboration is formed if all of the participating agents benefit from the collaboration. Therefore, a clean game requires that this game should be super-additive so that the revenue of each coalition is at least as much as the sum of any of its disjoint subsets. On the other hand, if the coalitional game consists of non-positive values only, it can be associated with a taxation scenario, in which larger coalitions induce higher taxes.

Definition 6 (easy-to-divide (ETD))**.** *A coalitional game* (N, v) *is easy-to-divide if all the agents that are not null-players are equivalent to each other.*

The intuition behind this definition is as follows. Let (N, v) be an easy-to-divide coalitional game, and let p be the number of null-players in (N, v). If we accept that a solution concept should follow the efficiency, null-player and symmetry axioms, then it is easy to calculate the allocation in an easy-to-divide game. Namely, all null-player agents receive a payoff of 0 and all of the other agents receive a payoff of $\frac{v(N)}{(|N|-p)}$. Clearly, this is also the Shapley value for this game.

Definition 7 (easy-to-explain (ETX))**.** *A coalitional game* (N, v) *is easy-to-explain if it is clean and easy-to-divide.*

Clearly, a game that is easy-to-explain represents a common scenario (since the game is clean) and it is easy to understand its payoff allocation (since the game is easy-to-divide).

Consider the following examples, which illustrate the ETX definition.

Example 1. *Let $N = \{a, b, c\}$. There are five games, (1)-(5), with the following characteristic functions:*

Coalition	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*
$\{a\}$	*1*	*0*	*1*	*0*	*1*
$\{b\}$	*0*	*0*	*2*	*0*	*1*
$\{c\}$	*0*	*0*	*0*	*0*	*0*
$\{a, b\}$	*1*	*-1*	*4*	*1*	*-1*
$\{a, c\}$	*1*	*0*	*2*	*1*	*1*
$\{b, c\}$	*0*	*0*	*2*	*0*	*1*
$\{a, b, c\}$	*1*	*-1*	*5*	*1*	*-1*

Games (1) and (2) are ETX. Indeed, it is natural to assume that a fair division of the revenue in game (1) assigns the total payoff to a, since b and c are null players. Similarly, a fair division of the tax in game (2) assigns -0.5 to the two equivalent agents a and b. Games (3) and (4) are clean but not ETD, and game (5) is ETD but not clean. Indeed, it is not straightforward to determine a fair division in these games.

Now, given an ETX game, it is possible to automatically generate a verbal explanation for the game based on the fact that the game is also ETD. Specifically, we need to find the equivalent agents and the null players. Then, it is easy to generate an explanation that points out which agents do not contribute to the outcome and which agents have an equal contribution and thus the total outcome should be equally divided between them. The explanation should also consider whether the game describes revenues or taxation. For example, if agents a and c are equivalent, agent b is a null-player, the game describes revenues, and the total revenue is $\$300$, it is possible to generate the following explanation:

> *"In this scenario, b does not contribute anything. a and c are identical and always contribute the same. Therefore, the total revenue, which is $\$300$, should be equally divided between a and b, and thus, the fair division is $a : \$150, b : \$0, c : \$150$."*

5 X-SHAP

In this subsection we propose the *X-SHAP* algorithm, which given any coalitional game, automatically decomposes the coalitional game into a number of ETX sub-games. Given the ETX sub-games, X-SHAP automatically generates verbal explanations for each of them (as described in Section 4) and presents the payoff allocations along with the explanations to human users. It is expected that humans will find the Shapley value payoff to be fair in each of the ETX sub-games, and thus the Shapley value for the given game, which is composed of the sub-games, should seem fair to humans as well.

The X-SHAP algorithm works as follows. It receives a coalitional game (N, v) as an input and provides a set X of characteristic functions that maintains the following two properties:

1. Each coalitional game (N, x), where $x \in X$, is easy-to-explain.
2. The sum of all the characteristic functions in X equals v. That is, $\sum_{x \in X} x = v$.

Note that since the Shapley value satisfies the additivity axiom, the sum of Shapley value payoffs assigned to each agent $i \in N$ in each characteristic function in X is equal to the Shapley value payoff

for i in (N, v). That is, $\forall i \in N, \sum_{x \in X} Sh_i(N, x) = Sh_i(N, v)$. Once the set X is generated, we generate verbal explanations for each of the sub-games.

Algorithm 1 describes the pseudo-code for X-SHAP. The algorithm iterates over all subsets $S \subseteq N$ in ascending order according to $|S|$. It maintains a characteristic function $accum$ that accumulates all the characteristic functions it builds in each iteration. For each subset S whose value in v is different from its value in $accum$, X-SHAP adds the following characteristic function x to X. For each subset of N, T, that contains S, $x(T)$ is set to the difference between $v(S)$ and $accum(S)$.

Algorithm 1: X-SHAP

Input : A coalitional game (N, v).
Output: A set of characteristic functions X, along with their verbal explanations.

```
X ← ∅
Let accum, x be characteristic functions on N
Initialize accum to 0 for any subset
for i ← 1 to |N| do
    for every S ⊆ N, such that |S| = i do
        Initialize x to 0 for any subset
        if v(S) ≠ accum(S) then
            for every T ⊇ S do
                x(T) ← v(S) − accum(S)
            X ← X ∪ {x}
            accum ← accum + x
Generate a verbal explanation for each x ∈ X
return X along with the verbal explanations
```

The number of characteristic functions in X is at most the number of subsets in N, which is, in fact, the size of the input. Denote $M = 2^{|N|}$ and $n = |N|$. A naive implementation of X-SHAP is $O(M^2)$. however, by using the following approach, we can reduce the complexity to $O(M^{\log_2 3}) \approx O(M^{1.58})$. For every subset $S \subseteq N$ we can get all its supersets $T \supseteq S$ by adding S to every subset of $N \setminus S$. Now, the number of subsets with i agents is $\binom{n}{i}$, and the number of supersets of every such subset is 2^{n-i}. Hence, the complexity of X-SHAP is:

$$\sum_{i=1}^{n} \binom{n}{i} 2^{n-i} = \sum_{i=1}^{n} \binom{n}{n-i} 2^{n-i} = \sum_{k=0}^{n-1} \binom{n}{k} 2^k \quad = \sum_{k=0}^{n} \binom{n}{k} 2^k 1^{n-k} - 2^n.$$

According to the binomial expansion formula, $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$, and thus,

$$\sum_{k=0}^{n} \binom{n}{k} 2^k 1^{n-k} - 2^n = (2+1)^n - M = 3^n - M = 3^{\log_2 M} - M \quad = O(M^{\log_2 3}).$$

Consider the following example, which illustrates the output that is generated by the X-SHAP algorithm.

Example 2. *Consider the following coalitional game (N, v), in which $N = \{a, b, c\}$, and v associates to every subset of N the following values:*

$\{a\}$	*0*
$\{b\}$	*0*
$\{c\}$	*100*
$\{a, b\}$	*300*
$\{a, c\}$	*200*
$\{b, c\}$	*100*
$\{a, b, c\}$	*500*

The Shapley payoff allocation for each of the agents in this game is $Sh_a(N, v) = 200, Sh_b(N, v) = 150$ *and* $Sh_c(N, v) = 150$*. It is not intuitive that this payoff allocation is indeed fair. For this input, X-SHAP outputs a set* X *with the following characteristic functions:*

Coalition	*(1)*	*(2)*	*(3)*
$\{a\}$	*0*	*0*	*0*
$\{b\}$	*0*	*0*	*0*
$\{c\}$	*100*	*0*	*0*
$\{a, b\}$	*0*	*300*	*0*
$\{a, c\}$	*100*	*0*	*100*
$\{b, c\}$	*100*	*0*	*0*
$\{a, b, c\}$	*100*	*300*	*100*

Each of these functions is ETX and their sum equals v*, i.e.,* $\sum_{x \in X} x = v$*. The Shapley payoff allocation for each of the coalitional games* (N, x)*, where* $x \in X$ *is:*

Agent	*(1)*	*(2)*	*(3)*
Sh_a	*0*	*150*	*50*
Sh_b	*0*	*150*	*0*
Sh_c	*100*	*0*	*50*

In addition, X-SHAP provides the following verbal explanation for each sub-game.

(1) "In this scenario, a *and* b *do not contribute anything. The entire revenue is contributed by* c *alone. Therefore, the total revenue, which is* $\$100$*, should solely go to* c*, and thus, the fair division is* $a : \$0, b : \$0, c : \$100$*."*

(2) "In this scenario, c *does not contribute anything.* a *and* b *are identical and always contribute the same. Therefore, the total revenue, which is* $\$300$*, should be equally divided between* a *and* b*, and thus, the fair division is* $a : \$150, b : \$150, c : \$0$*."*

(3) "In this scenario, b *does not contribute anything.* a *and* c *are identical and always contribute the same. Therefore, the total revenue, which is* $\$100$*, should be equally divided between* a *and* c*, and thus, the fair division is* $a : \$50, b : \$0, c : \$50$*."*

Given these payoff allocations and their verbal explanations, it is quite likely that human users will accept each of them as fair. The sum of all the payoff allocations of each agent is indeed equal to the shapely value of the original game (N, v)*, i.e.* $\forall i \in N, \sum_{x \in X} Sh_i(N, x) = Sh_i(N, v)$*.*

X-SHAP Properties

We now prove that the set X of characteristic functions that is returned by Algorithm 1 maintains the required properties.

Theorem 1. *Each coalitional game* (N, x)*, where* $x \in X$*, is easy-to-explain.*

Proof. Given a characteristic function $x \in X$, it corresponds to a subset $S \subseteq N$. X-SHAP constructs x such that it assigns a non-zero value, val, for every $T \supseteq S$, and a zero value otherwise. Therefore, for any agent $i \notin S$ and for every subset $P \subseteq N \setminus \{i\}$, it holds that $x(P \cup \{i\}) = x(P)$. That is, every agent $i \notin S$ is a null player. On the other hand, every agent $i \in S$ is not a null player, since $x(S \setminus \{i\}) = 0$ but $x(S) = val \neq 0$. In addition, for every two agents $i, j \in S$ and any subset $P \subseteq N \setminus \{i, j\}$, it holds that $x(P \cup \{i\}) = x(P \cup \{j\})$. That is, every two agents $i, j \in S$ are equivalent. Therefore, the coalitional game (N, x) is ETD. Finally, for every pair of disjoint subsets P_1, P_2, these are the possible cases:

- $P_1, P_2 \not\supseteq S$, and thus $v(P_1) = v(P_2) = 0$. Now, if val is positive then $v(P_1 \cup P_2) \geq v(P_1) + v(P_2)$, and if val is negative then $v(P_1 \cup P_2) \leq v(P1) + v(P2)$.
- Without loss of generality, $P_1 \supseteq S$ but $P_2 \not\supseteq S$. We get that $v(P_1) = val$ but $v(P_2) = 0$. In addition, since $P_1 \cup P_2 \supseteq S$, $v(P_1 \cup P_2) = val = v(P_1) + v(P_2)$.

Therefore, if val is positive then x is super-additive and if val is negative then x is sub-additive. That is, (N, x) is clean, and since (N, x) is also ETD it is ETX. □

Theorem 2. *The sum of all the characteristic functions in X equals v. That is, $\sum_{x \in X} x = v$.*

Proof. The algorithm iterates over all $S \subseteq N$. At the end of an iteration in which $S \subseteq N$ is considered, $accum(S)$ equals $v(S)$. This is because either $accum(S)$ already equals $v(S)$ or $x(S)$ is set to $v(S) - accum(S)$ in line 9, and after line 11 $accum(S)$ becomes $v(S)$. After considering S the algorithm does not consider any $S' \subseteq S$, and thus all following iterations do not change $accum(S)$. Finally, according to the algorithm construction, $accum$ holds the sum of all functions $x \in X$ when the algorithm terminates. Hence, $\sum_{x \in X} x = accum = v$. □

We note that a characteristic function $x \in X$, that correspond to some coalition $S \subseteq N$, may contain negative values even if v consists of only non-negative values. This situation will occur when the sum of all the characteristic functions constructed before x is higher than $v(S)$. We show that any procedure that decomposes a coalitional game with a non-negative characteristic function into a number of ETX sub-games, cannot avoid using sub-games with a negative characteristic function.

Theorem 3. *There exist coalitional games with non-negative characteristic functions such that any decomposition into ETX sub-games results in at least one sub-game with negative characteristic function.*

Proof. Consider the following coalitional game (N, v), which is the classical "glove game", in which $N = \{a, b, c\}$ and for every $S \subseteq N$,

$$v(S) = \begin{cases} 1 & S \in \{\{a, b\}, \{a, c\}, \{a, b, c\}\} \\ 0 & else. \end{cases}$$

Assume towards contradiction that (N, v) can be decomposed into ETX sub-games, such that none of their characteristic functions consist of negative values. Let X be the set of these characteristic functions, and let $X_S^+ \subseteq X$, where $S \subseteq N$, be the set of all the characteristic functions in X that assign a value greater than 0 for the coalition S. That is, for each $x \in X_S^+$, $x(S) > 0$. Since $\sum_{x \in X} x(\{a, b\}) = v(\{a, b\}) = 1$, and every $x \in X$ does not consist of negative values, it should hold that $\sum_{x \in X^+_{\{a,b\}}} x(\{a, b\}) = 1$. Since each $x \in X^+_{\{a,b\}}$ does not consist of negative values and each sub-game is clean, then by definition x is super-additive; therefore, $\sum_{x \in X^+_{\{a,b\}}} x(\{a, b, c\}) \geq 1$. Furthermore, since $v(\{a, b, c\}) = 1$ and x is non-negative, it must hold that $\sum_{x \in X^+_{\{a,b\}}} x(\{a, b, c\}) = 1$. Similarly, for the set $X^+_{\{a,c\}}$, $\sum_{x \in X^+_{\{a,c\}}} x(\{a, b, c\}) = 1$.

Now, for any $x \in X^+_{\{a,b\}}$ it must hold that $x \in X^+_{\{a,c\}}$, otherwise, if there is $x' \in X^+_{\{a,b\}}$ such that $x'(\{a,c\}) = 0$ then $\sum_{x \in X^+_{\{a,c\}}} x(\{a,b,c\}) + x'(\{a,b,c\}) > 1$. Finally, since $v(\{a\}) = v(\{b\}) = v(\{b,c\}) = 0$ and every $x \in X^+_{\{a,b\}}$ is non-negative, $x(\{a\}) = x(\{b\}) = x(\{b,c\}) = 0$. However, $x(\{a,b\}) > 0$ and thus a and b are not null players in the sub-game (N,x), but $x(\{c\} \cup \{a\}) = x(\{a,c\}) > 0$ and $x(\{c\} \cup \{b\}) = x(\{b,c\}) = 0$. That is, a and b are not equivalent and thus the sub-game (N,x) is not ETX, which is a contradiction. □

6 Experimental Design

We begin our evaluation by validating the concept of ETX. To that end, we first ran a survey on Mechanical Turk [20]. The participants were first given an appropriate background on coalitional games in general and instructions specific to the survey. To verify that the participants read and understood the instructions, each participant was required to correctly answer a short quiz with four questions in order to proceed. The participants were then presented with an ETX game in which the agents were referred to as entities, and the values of the characteristic function were referred to as revenues. The game was composed of three entities, marked as a, b, c, and the participants were presented with a table of revenues of the entities when they are alone and when they collaborate with each other. Then, each participant was presented with one of the following possible payoff allocation: 1. The Shapley payoff allocation. 2. The inverse allocation. In this allocation, the agents that are null players equally share the total revenue and all the other agents receive a payoff of $\$0$.

The participants were asked to rate the proposed payoff allocation by indicating to what extent they agree or disagree that it is fair. The participants could choose one of seven options on a Likert scale [15], between "strongly agree" (7) to "strongly disagree" (1), with the middle being "neither agree nor disagree" (4). Likert scale is commonly used in research and surveys to measure attitude, providing a range of responses to a given question or statement. We used two ETX games: the first two ETX games in the set X from Example 2 (as shown in columns (1) and (2) there). Each payoff allocation was presented to 30 different participants for each of the two ETX games. Overall, we had 120 participants in this initial experiment, and the reward for each participant was $\$0.3$.

In our main experiment, we evaluated the explanation generated by X-SHAP. We ran a similar survey on Mechanical Turk, in which the participants were presented with a coalitional game and the Shapley payoff allocation. Then, each participant was presented with one of the following:

1. X-SHAP's complete explanation. Participants were able to switch between all the sub-games so that they could examine each sub-game individually. For each sub-game they were presented with its allocation according to the Shapley value with a brief verbal explanation. Finally, as shown in Figure 1, each participant was shown how the sum of all the sub-games and their Shapley value allocation equal to the original game and its Shapley value.

2. X-SHAP's decomposition into sub-games without their verbal explanations.

3. Sub-game decomposition: A heuristic decomposition of the game into sub-games, so that for each subset whose value in the original game is different from 0, there is a sub-game where this subset gets its original value and all other subsets get the value 0. The graphical user interface was identical to that of X-SHAP's.

4. Marginal contribution: A verbal explanation describing the largest marginal contribution of each agent

5. Fixed: A fixed general explanation of the Shapley value that was taken from the *Spliddit* website [8]; it states that the allocation is based on the marginal contribution of each agent to each possible coalition.

6. No explanation at all.

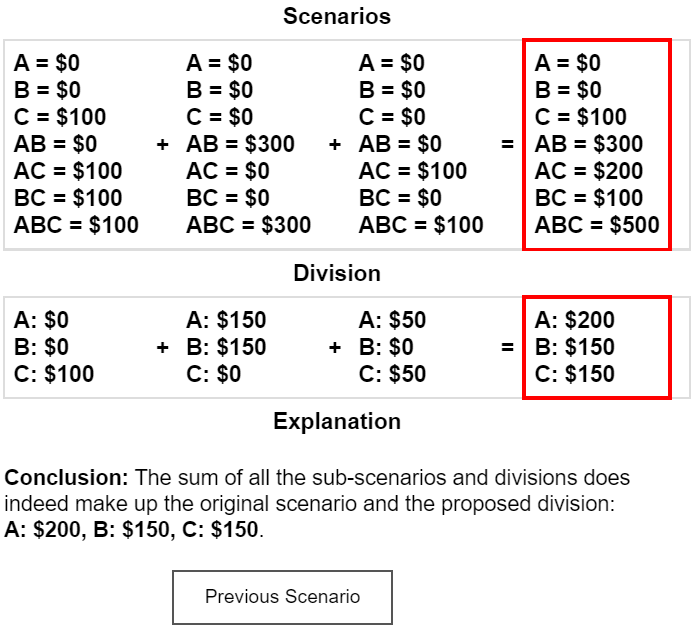

Figure 1: Screenshot from the survey of the X-SHAP explanation.

The participants were asked to rate the proposed payoff allocation on a Likert scale, as in the initial experiment. The participants were then presented with a different coalitional game along with its Shapley payoff allocation accompanied by one of the above-mentioned explanation types (different from the explanation type received for the first scenario). They were again asked to rate the proposed payoff allocation.

Table 1 specifies the coalitional games that we used for the survey. In each of these games (N, v), $N = \{a, b, c\}$, and all revenues are non-negative. The Shapley payoff allocation for each of the scenarios appears in Table 2.

Table 1: The coalitional games that we used for the survey.

Coalition	(1)	(2)	(3)	(4)	(5)	(6)
$\{a\}$	200	0	0	0	100	300
$\{b\}$	200	100	0	0	200	0
$\{c\}$	100	200	100	0	300	500
$\{a, b\}$	400	300	300	300	200	500
$\{a, c\}$	600	400	200	300	300	100
$\{b, c\}$	600	300	100	0	300	200
$\{a, b, c\}$	800	700	500	300	350	600

We chose these coalitional games as they represent a variety of scenarios: in game (1) all the values are greater than zero, and agents a and b are equivalent. In game (2) the value of $\{a\}$ is zero and a and b are not equivalent, but the Shapley payoff for a and b is nevertheless identical. In game (3) the value of $\{a\}$ and $\{b\}$ is zero, there are no equivalent agents, but the Shapley payoff for b and c is nevertheless identical. In game (4) the value of $\{a\}$, $\{b\}$ and $\{c\}$ is zero, yet only b and c are

Table 2: The Shapley payoff allocation for each of the scenarios from Table 1.

Agent	(1)	(2)	(3)	(4)	(5)	(6)
Sh_a	250	200	200	200	50	250
Sh_b	250	200	150	50	100	150
Sh_c	300	300	150	50	200	200

equivalent agents. Note also that game (4) is the glove game mentioned above. The characteristic functions in games (1)-(4) are super-additive. This attribute is common since if two (or more) agents collaborate, they are expected to gain more than each would have gained by herself. Yet, we also tested two less common scenarios: In game (5) the characteristic function is sub-additive, while in game (6) the characteristic function is neither super-additive nor sub-additive.

Each of the six explanation types was presented to 35 different participants for each of the six scenarios. Overall, we had 630 participants in the main experiment, each answering two different scenarios with different explanations. The reward for each participant was $\$0.5$. In total, i.e., in both the initial and the main experiments, the average age of the participants was 39 with 453 males and 284 females; 13 participants chose not to specify their gender. We set a requirement on Mechanical Turk that the approval rate of the works must be at least 99% and did not require the Turkers to be masters.

7 Results

In our initial experiment, we validate the concept of ETX with two ETX games. The average fairness rating of the Shapley allocation was 5.76 and 5.83, which is significantly greater ($p < 0.0001$) than with the inverse allocations, in which it was only 2.5 and 2.13. This validates our assumption that Shapley allocation for ETX games are perceived as fair. We use these results as an indication of the maximum and minimum average fairness rating that can be obtained in our setting.

In our main experiment, we evaluate the explanations generated by X-SHAP. The results, presented in Figure 2, were obtained by averaging over the ratings of the participants. As depicted by the figure, the explanations that were generated by X-SHAP (with or without the verbal explanations) outperformed all other explanations in terms of fairness rating in all the scenarios examined. That is, the human participants perceive the payoff allocation fairer if they receive the explanations that are generated by X-SHAP. Overall, the average fairness rating in scenarios in which the X-SHAP explanation was provided is 5.32, which is very close to the average rating of the Shapley allocation in the ETX games. We note that other explanation-types occasionally obtained low ratings, which indicate that the Shapley allocation may be perceived as unfair.

For checking the statistical significance, we ran repeated measures ANOVA test, which considers both the scenario and the type of explanation. The ANOVA test lead to statistically significant differences ($p < 0.01$) between X-SHAP and all other types of explanations (except X-SHAP without verbal explanations). Indeed, analyzing the outcomes of the Likert scale, and the use of parametric tests to analyze ordinal data in general, has been subject to an active and ongoing debate [19]. We thus conducted also a non-parametric test, an ordinal logistic regression analysis, which is used to assess the difference between two methods with ordinal values, such as ratings and pain level reporting [10]. The ordinal logistic regression analysis also demonstrated the significance of the results.

We note that the explanations that were generated by X-SHAP for scenarios (4)-(6) yielded a lower average of fairness rating compared to the explanations for scenarios (1)-(3). A possible reason is that while scenarios (1)-(3) include only characteristic functions with positive values, in

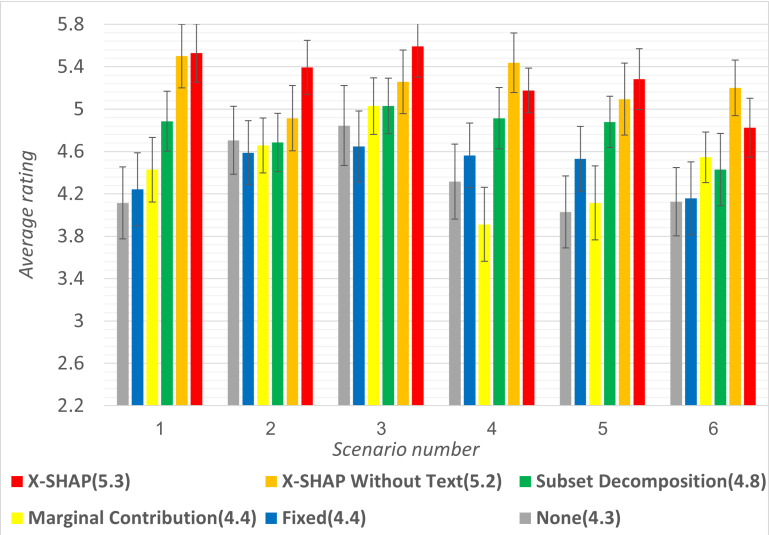

Figure 2: Average rating of fairness for all explanation types in each of the six scenarios. Error bars present the standard error. The average appears in parentheses.

scenarios (4)-(6) the explanations include characteristic functions with positive values along with characteristic functions with negative values. The combination of positive and negative characteristic functions in one explanation may be confusing. However, this phenomenon cannot be avoided according to theorem 3. Nevertheless, the interaction effect between the scenario and the type of explanation is non-statistically significant. We also note that Scenario (6) has the lowest average fairness rating for X-SHAP. A possible reason is that its characteristic function is neither super-additive nor sub-additive, and thus, represents a less intuitive scenario.

8 Conclusions and Future Work

The Shapley value is termed the most important normative division scheme in cooperative game theory. However, in some scenarios, its payoff allocation may seem unfair to humans. We provided the first automatic method that generates customized explanations for the Shapley value. Our approach does not directly use psychological insights regarding the perception of fairness by humans. Instead, we utilize known mathematical axioms, and show that they can be used for increasing the rating of fairness of the Shapley allocation.

Recall that the number of sub-games that X-SHAP shows to the user depends on the scenario and the number of agents. Therefore, in games with many agents, X-SHAP may be required to present its users with hundreds of sub-games, each game consisting of all subsets of the agents. In future work, we intend to address this issue and propose three complementary approaches.

First, instead of presenting all the coalitions of a sub-game, X-SHAP can alternatively state that a specific coalition and any coalition containing it receive some payoff. Furthermore, instead of presenting all sub-games, X-SHAP can present for a user only the sub-games in which she receives a non-zero payoff. Moreover, X-SHAP can present the explanations in an interactive process, in which a user is provided with evidences (i.e., sub-games) until she is convinced that the provided allocation is fair. This interactive process requires presenting the stronger evidence earlier during the process; this raises several interesting questions related to human perception of fairness. Alternatively, the sub-games can be provided only upon requests from the user. That is, the user will ask to see all sub-games where the payoff for a specific agent or coalition is not zero.

Acknowledgment

This research was supported in part by the Ministry of Science, Technology & Space, Israel.

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[2] O. Cailloux and U. Endriss. Arguing about voting rules. In *AAMAS*, 2016.

[3] Mark G. Core, H. Chad Lane, Michael van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. Building explainable artificial intelligence systems. In *AAAI Conference on Artificial Intelligence*, 2006.

[4] Geoffroy de Clippel. Membership separability: A new axiomatization of the shapley value. *Games and Economic Behavior*, 108:125–129, 2018.

[5] Geoffroy de Clippel and Kareen Rozen. Fairness through the lens of cooperative game theory: An experimental approach. Technical Report 1925, Cowles Foundation for Research in Economics, Yale University, 2013.

[6] Greg d'Eon and Kate Larson. Testing axioms against human reward divisions in cooperative games. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, volume 19, pages 312–320, 2020.

[7] Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable ai really mean? a new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*, 2017.

[8] Jonathan Goldman and Ariel D. Procaccia. Spliddit: Unleashing fair division algorithms. *SIGecom Exch.*, 13(2):41–46, January 2015.

[9] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4(37), 2019.

[10] Frank E Harrell. Ordinal logistic regression. In *Regression modeling strategies*, pages 311–325. Springer, 2015.

[11] John C. Harsanyi. A bargaining model for the cooperative n-person game. In Albert William Tucker and Robert Duncan Luce, editors, *Contributions to the Theory of Games (AM-40), Volume IV*, pages 325–356. Princeton University Press, 1959.

[12] John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.

[13] Sergiu Hart. Shapley value. In *Game Theory*, pages 210–216. Springer, 1989.

[14] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable ai systems for the medical domain?, 2017.

[15] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396, 2015.

[16] Sarit Kraus, Amos Azaria, Jelena Fiosina, Maike Greve, Noam Hazon, Lutz Kolbe, Tim-Benjamin Lembcke, Jorg P Muller, Soren Schleibaum, and Mark Vollrath. Ai for explaining decisions in multi-agent environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13534–13538, 2020.

[17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, volume 31, pages 4768–4777, 2017.

[18] Hervé Moulin. *Fair division and collective welfare*. MIT press, 2004.

[19] Yazan Mualla, Igor Tchappi, Timotheus Kampik, Amro Najjar, Davide Calvaresi, Abdeljalil Abbas-Turki, Stéphane Galland, and Christophe Nicolle. The quest of parsimonious xai: a human-agent architecture for explanation formulation. *Artificial Intelligence*, 302:1–26, 2021.

[20] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[21] Dominik Peters, Ariel D Procaccia, Alexandros Psomas, and Zixin Zhou. Explainable voting. In *NeurIPS*, volume 33, pages 1525–1534. Curran Associates, Inc., 2020.

[22] Ariel D Procaccia. Axioms should explain solutions. In *The Future of Economic Design*, pages 195–199. Springer, 2019.

[23] D. Schmeidler. The nucleolus of a characteristic function game. *Siam Journal on Applied Mathematics*, 17:1163–1170, 1969.

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

[25] Ari Stern and Alexander Tettenhorst. Hodge decomposition and the shapley value of a cooperative game. *Games and Economic Behavior*, 113:186–198, 2019.

[26] Eyal Winter. The Shapley value. *Handbook of game theory with economic applications*, 3: 2025–2054, 2002.

Contact Information

Meir Nizri
Ariel University, Israel
Email: meir.nizri@msmail.ariel.ac.il

Amos Azaria
Ariel University, Israel
Email: amos.azaria@ariel.ac.il

Noam Hazon
Ariel University, Israel
Email: noamh@ariel.ac.il