ARIEL UNIVERSITY

MASTER THESIS

Explaining Fair Allocations and Recommendations

Author:
Meir Nizri

Supervisors:
Prof. Amos Azaria
Dr. Noam Hazon

Department of Computer Science

October 31, 2022

ARIEL UNIVERSITY

MASTER THESIS

Explaining Fair Allocations and Recommendations

Author:
Meir Nizri

Supervisors:
Prof. Amos Azaria
Dr. Noam Hazon

Department of Computer Science

October 31, 2022

Acknowledgements

Ariel University

Abstract

Faculty of Natural Sciences

Department of Computer Science

Master

Explaining Fair Allocations and Recommendations

by Meir Nizri

As artificial intelligence (AI) becomes more advanced and significant in extensive aspects of our daily lives, humans are challenged to comprehend and retrace how the algorithm came to a result. There are many advantages to understanding how an AI-enabled system has led to a specific output. Explainability allows those affected by a decision to challenge or change that outcome, it might also be necessary to meet regulatory standards, or it can help developers to ensure that the system is working as expected. In this thesis we developed methods for explaining artificial intelligence algorithms in two common domains: game theory and recommendation systems.

In the first chapter of this thesis, we explain allocations according to the Shapley value. The Shapley value is one of the most important normative division schemes in cooperative game theory, satisfying basic axioms. However, some allocations according to the Shapley value may seem unfair to humans. 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.

In the second chapter of this thesis, we explain recommendation systems output. Recommendation systems are widely used and are present in many applications, such as movie recommendation, product sales, and content providers. However, current recommendation systems are usually stern and lack the ability to explain their decisions or allow the users to question them. We develop an automatic method that generates contrastive explanations for recommendation systems based on items features and users' preferences. That is, once receiving a recommendation (e.g., to buy a Samsung S22), the users will have the option to ask the system why it did not recommend a specific different item (e.g., Xiaomi 12). With our method the system has the ability to reply with a meaningful and convincing personalized explanation

(e.g., it might seem that a good camera is very important to a specific user, and the Samsung S22 includes a better camera than the Xiaomi 12).

In both studies, we run experiments with human participants and show that when applying our methods, humans are more convinced that the output of the AI is better for them than when not using any explanation or using explanations generated by other methods.

Contents

Chapter 1

Improving the Perception of Fairness in Shapley-Based Allocations

1.1 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 [1, 2] or the nucleolus [3], but the Shapley value [4] has been termed the most important normative division scheme in cooperative game theory [5]. 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 [6] (see Section 1.3.1 for formal definitions). These axioms admit strong normative and positive interpretations [7]. We note that there are several equivalent sets of axioms that characterize the Shapley value [8].

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 1.4.2). 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 thesis, 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" [9], and this direction has been successfully applied in the field of fair division by the *Spliddit* website [1] [10]. 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 [4].

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

[1] http://www.spliddit.org/

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 the first part of this thesis is that it provides the first successful automatic method that generates customized explanations of the Shapley allocation for any given coalitional game.

We published and presented a paper based on this thesis at the 2022 Cognitive Science Conference [11].

1.2 Related Works

Our work is related to the field of Explainable AI (XAI) [12, 13]. 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 [14]. Other XAI systems are designed to provide explanations, comprehensible by humans, for legal or ethical reasons [15]. For example, an AI system for the medical domain might be required to explain its choice for recommending the prescription of a specific drug [16]. Indeed, most of the work on XAI concerned the explanation of a machine learning based model. In this thesis, 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 [17], explainable decisions in multiagent environments.

The work that is closest to ours, albeit in the context of voting, is by Cailloux and Endriss [18]. 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. [19], 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 [20]. 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 [7] 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 [10] 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 (SHapley Additive exPlanations) [21]. 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.

1.3 Fundamentals of the X-SHAP Approach

1.3.1 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., $v : 2^N \to \mathbb{R}$. 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 [4]. Formally, for each agent i,

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

1.3.2 Shapley Value Axioms

The Shapley value is the only solution concept that simultaneously satisfies the following axioms [6].

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, $v+w : 2^N \to \mathbb{R}$, 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 $\phi_i(N,v+w) = \phi_i(N,v)+\phi_i(N,w)$.*

1.3.3 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* $a : \$150, b : \$0, c : \$150$*."*

1.3.4 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 subgames. Given the ETX sub-games, X-SHAP automatically generates verbal explanations for each of them (as described in Section 1.3.3) 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 subgames.

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 subgame.

(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.* $\forall i \in N, \sum_{x \in X} Sh_i(N,x) = Sh_i(N,v)$.

1.3.5 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. □

1.4 Experimental Evaluation

1.4.1 Experimental Design

We begin our evaluation by validating the concept of ETX. To that end, we first ran a survey on Mechanical Turk [22]. 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 [23], 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 1.1, 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.

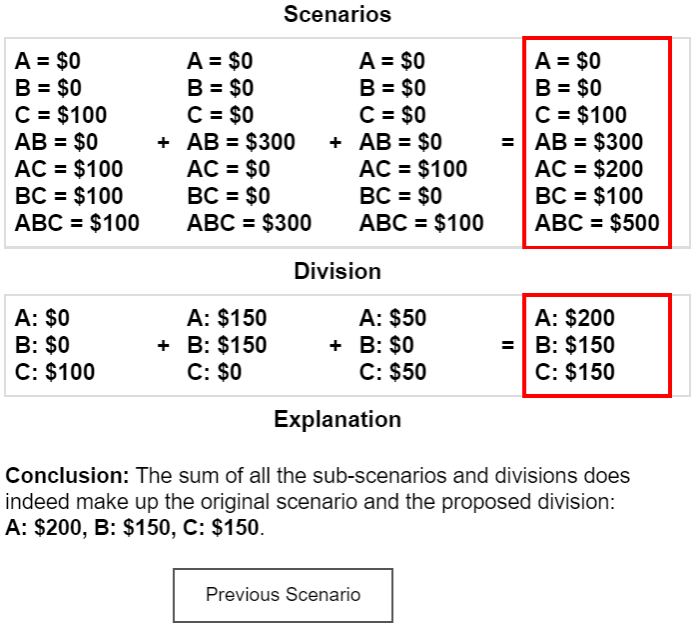

FIGURE 1.1: Screenshot from the survey of the X-SHAP explanation.

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. For example, such decomposition for the coalitional game in Example 2 would include the sub-games in the following table.

Coalition	(1)	(2)	(3)	(4)	(5)
$\{a\}$	0	0	0	0	0
$\{b\}$	0	0	0	0	0
$\{c\}$	100	0	0	0	0
$\{a,b\}$	0	300	0	0	0
$\{a,c\}$	0	0	200	0	0
$\{b,c\}$	0	0	0	100	0
$\{a,b,c\}$	0	0	0	0	500

4. Marginal contribution: A verbal explanation describing the largest marginal contribution of each agent. For example, this explanation for the coalitional game in Example 2 would be as follows:

   > *"The idea behind this division is that when a is added to b and c, it adds a contribution of* \$400*, when b is added to a, it adds a contribution of* \$300*, and when c is added to a, it adds a contribution of* \$200*."*

5. Fixed: A fixed general explanation of the Shapley value that was taken from the *Spliddit* website [10]; it states that the allocation is based on the marginal contribution of each agent to each possible coalition.
6. No explanation at all.

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 1.1 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 1.2.

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 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

TABLE 1.1: 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

TABLE 1.2: The Shapley payoff allocation for each of the scenarios from Table 1.1.

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

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.

1.4.2 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 1.2, 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.

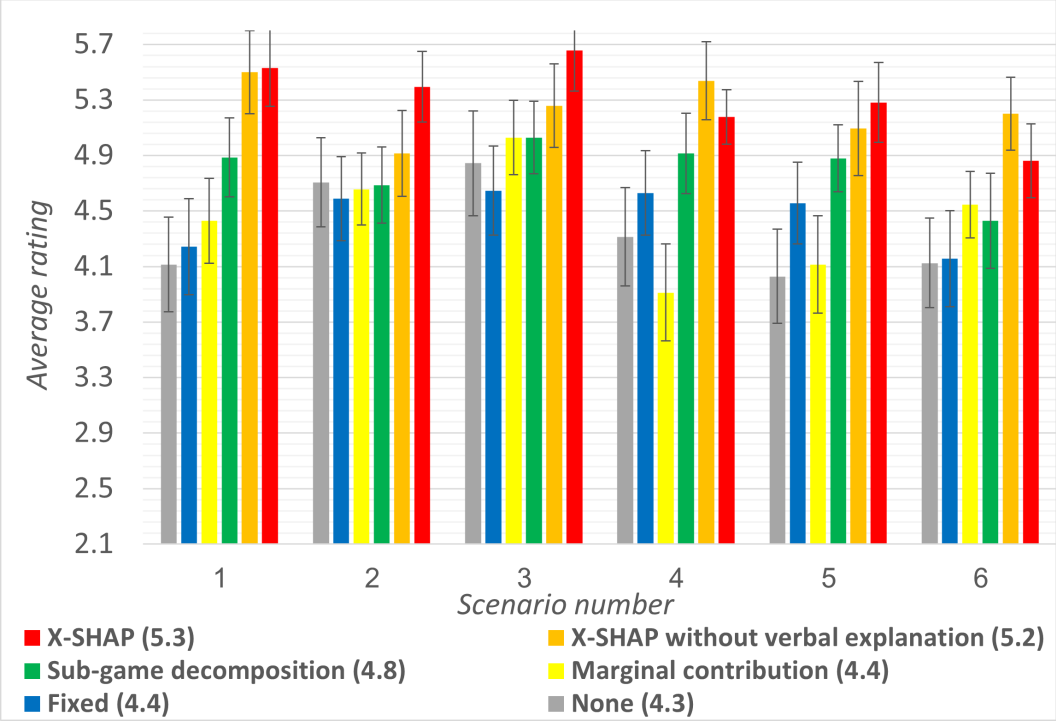

FIGURE 1.2: Average rating of fairness for all explanation types in each of the six scenarios. Error bars present the standard error. The average appears in parentheses.

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 [24]. 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 [25]. 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 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.

1.5 Future Work

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 different 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.

Chapter 2

Contrastive Explanations for Recommendation Systems

2.1 Introduction

Recommendation systems are becoming significant in extensive aspects of our daily lives. In some areas, such as medicine [26] and finance [27], recommendation systems are used as tools that help making decisions at high stakes. In these settings, the system is expected to justify its recommendations, since the cost of a wrong decision may be high. However, state-of-art techniques used in recommendation systems are becoming more and more inscrutable due to the complexity of models and high dimension of data on which the models are trained. Arguably, the most significant means to increase trust and transparency in recommendation systems is by providing explanations for the system's decisions that are concise and understandable to humans [28].

In order to explain the outputs of a recommendation system, two strategies that are common in the field of explainable AI can be applied. The first is to adopt interpretable models, whose decision mechanism is transparent, and thus, we can naturally provide explanations for the model decisions [29, 30]. Explanations of this type are called ante-hoc, and they seek to understand the inner workings of a model while the model is in the process of making decisions. However, they often suffer from lower prediction accuracy, which is known as the trade-off between complexity and accuracy [31]. In addition, when recommendations are provided as an external service, the inner workings of the system are the intellectual property of the service provider and cannot be revealed. Therefore, in this work we use the second strategy, which is to provide post-hoc explanations, i.e., after the recommendation has been given. Post-hoc explanations do not precisely reflect the underlying recommendation model. Instead, they present rationale, plausible, and valuable information to the user. Consequently, post-hoc explanations are independent of the main recommendation algorithm, and therefore they have the potential to be used across a greater diversity of recommendation systems.

Indeed, what constitutes good explanations has been extensively studied in social science, cognitive science, and psychology [32, 33]. In particular, P. Lipton [34] investigated different types of explanations, and concluded that they should be contrastive. That is, instead of explaining why an event happened, explanations are more

convincing if the person receiving the explanation can ask why one event happened rather than another specific event. We follow this idea, and develop a method that generates contrastive explanations for recommendation systems. A contrastive explanation in recommendation system allow the user to pose questions of the form "why did you recommend this item rather than another specific item?". With our method the system can reply with a meaningful and convincing personalized explanation which is based on the features of the recommended item, the contrasting item or the user's preferences. To the best of our knowledge, currently, no recommendation system allows users to pose contrastive questions.

Given a user who was recommended an item p, the user may ask the system why it did not recommend another item q. The essence of our explanation is to select a set of features and show the user the differences between p and q in the selected features. For example, suppose that the item p is the Galaxy S22 cell phone and the item q is the iPhone 13 cell phone. If the selected set of features is price and main camera resolution, then the generated explanation should be "The recommended cell phone Galaxy S22 costs \$528, compared to iPhone 13, which costs \$699; The recommended cell phone Galaxy S22 main camera's resolution is 50 MP, compared to a resolution of 12 MP in the iPhone 13".

In order to select meaningful and convincing set of features, we develop an algorithm, which we termed *CX-RS*. CX-RS measures the influence of each item feature by sampling hypothetical items that are close (in terms of their feature values) to p, q, and in the area between them. The new items are used to train an interpretable model (e.g., linear regression, decision tree) from which we derive weights that represent the importance of each feature in the area between p and q. We then iterate over the features, and select only the first set of features such that if we replace their values in q to their values in p, the recommendation system will rank q as equal or higher than p.

In order to evaluate the performance of CX-RS, we conducted a survey with human participants. The data we used in the survey is a cell phone dataset. Each participant in the survey was presented with 10 randomly selected cell phones with all the necessary features to evaluate them. The participant was asked to rate each cell phones and provide demographic information. Then, two cell phones were presented. The first is the most recommended cell phone for the participant, selected by a state-of-the-art recommendation system. The participant was told that the second cell phone was preferred by a user with preferences similar to his. Each participant was then presented with four explanations why the recommended cell phones is better for him. One of the explanations was generated by CX-RS and the rest served as baselines. Finally, the participant was asked to rate each explanations by indicating to what extent the explanation convinces him that the recommended cell phone is indeed better for him. Overall, 100 different people participated in the survey. The explanations that were generated by CX-RS achieved higher average rating compared to the other explanations.

To summarize, the main contribution of the second part of this thesis is that it

provides a method that can generate more human-like explanation to recommendation systems and can be applied to any feature-based recommendation system models. More specifically, we propose a method that generates post-hoc contrastive explanation for recommendation systems.

2.2 Related Work

Explanation methods in AI can be classified into local explanation methods and global explanation methods [35]. Global explanation methods can explain the entire model behavior, how each feature can influence the results of the model. On the contrary, local explanation methods zoom in on a single instance to examine how the output has been generated. Local explanation offering more personalised explanations, and therefore we choose this method in our work.

Based on the model adopted in the recommendation system, explanations can include different types of information to explain the output of the recommendation system. Similar user/item style explanations [36], feature-based explanations [37], social explanations [38], and context-aware explanations [39]. In our explanation, we focus on items features since they can show the difference between two items in the most simplest and yet meaningful way.

There are several types of post-hoc explanations in recommendation systems. One common type provides explanations that answer why a specific item was recommended. Specifically, Peake and Wang [40] presented a post-hoc approach that extracts explanations from latent factor-based recommendation systems, by training association rules on the output of a matrix factorization black-box model. Nobrega and Marinho [41] introduced LIME-RS, which was later improved by Chanson et al [42]. LIME-RS is an adaptation of LIME to recommendation systems. LIME is a popular approach for explaining a machine learning model's output by identifying the top-n features that have the greatest impact on the model's output [43]. SHAP is another common approach for explaining a machine learning model's output; it relies on computing the average, over all permutations, of the marginal contributions of each feature [21]. Indeed, Guo et al. [44] adapt SHAP to derive post-hoc explanations for recommendation systems. Shmaryahu et al [45] use a set of simple, easily explainable recommendation algorithms to provide post-hoc explanations for a complex recommendation system. Their method attempts to find a simple explainable recommendation system that agrees with the complex model on a recommended item and applies the explanation of the simple model.

Another type of explanations is counterfactual explanations, which consider changes to features and events that alter the outputs of the recommendation system [46]. A typical counterfactual explanation describes a causal situation: "If the specific event did not occur, a different item would have been recommended". Unlike methods that try to approximate the original models, counterfactual explanations examine changes in the output of the original model and therefore have high fidelity [47]. In the recommendation system domain, there exists some works that aim to provide counterfactual explanations to explain recommendations. Ghazimatin et al.

[48] introduce PRINCE, which provides an explanation by finding a set of minimal actions performed by the user that, if removed, changes the recommendation to a different item. Given a recommendation, PRINCE uses a polynomial-time optimal algorithm for finding this minimal set of a user's actions from an exponential search space, based on random walks over dynamic graphs. This approach was later improved by Kaffes et al. [47], who used normalized length and the importance of a candidate to guide the search. Zhong and Negre [49] provides counterfactual explanation by sorting all features according to their SHAP values, and selects the minimum of the first features that, if we change their value to another random value, the recommendation system will change its recommendation.

Our proposed work should not be confused with interactive recommendation systems [50, 51], a field that focuses on updating the system-model based on interaction with the user and does not consider interactive explanations as we do.

2.3 CX-RS

Given a user, we assume that the goal of a recommendation system is to recommend an item that best fits the user preferences. Thus, the recommendation system rates each item and the item with the highest rating is recommended to the user. In our setting, the recommendation system uses a set of item features F, for example the item's brand or price; the features can be numeric or categorical. The recommendation system is not limited to using the item features and may also use features of the user (e.g., demographic information) and additional data (e.g., collaborative-based).

We developed CX-RS, an algorithm that generate post-hoc contrastive explanations that be applied to any feature-based recommendation system models. Given a user who was recommended an item p, the user may ask the system why it did not recommend another item q, and the system must be able to reply with a convincing explanation that is suited to the user's preferences. We propose to select a set of features in order to generate an explanation. The essence of the explanation is to show the user the differences between p and q in the selected features.

In CX-RS algorithm we propose a sampling-based approach, which measures the influence of each feature by sampling hypothetical items that are close (in terms of their feature values) to p, q, and in the area between them. We first generate new items. For each new item, we independently sample each of its features from a uniform distribution between p's value for that feature and q's value for it. That is, for every new item x, each feature $f \in F$ of x, denoted by x_f, is independently sampled from $U\{\min\{p_f, q_f\}, \max\{p_f, q_f\}\}$. In addition, the rating of the recommendation system for each new item is computed. The new items are used to train an interpretable model (e.g., linear regression, decision tree). The model is used to derive weights, which their absolute values represent the importance of a unit of each feature in the area between p and q. We multiply these weights, for each feature f, by the distance $p_f - q_f$, which will provide the explanatory power of each feature.

We then iterate over the features, where they are ordered by the explanatory powers (in descending order). We select only the first set of features such that if we replace their values in q to their values in p, the recommendation system will rank

q as equal or higher than p (see Algorithm 2). Intuitively, this approach finds the disadvantages of q such that if they are addressed (using the values of the features in p), q will be ranked as equal or higher than p. We consider the opposite approach, i.e., replacing the values of features in p with their values in q. Intuitively, this approach finds the advantages of p. We also consider limiting the number of selected features by some constant, k. As mentioned before, the explanation is generated from this set of features.

Algorithm 2: CX-RS

Input : Recommendation system R; Set of item features F; User u; Recommended item p; Contrastive item q; Number of samples n.
Output: A set of explanatory features.

$X, E \leftarrow \emptyset$
for $i \leftarrow 1$ *to* n **do**
  for $f \in F$ **do**
    $x_f \leftarrow$ sample from $U\{\min\{p_f, q_f\}, \max\{p_f, q_f\}\}$
  $X \leftarrow X \cup x$
$Y \leftarrow R(X, u)$
train an interpretable RS, T, on (X, Y)
$W \leftarrow T$'s weights multiplied by $(p - q)$
for $f \in F$ *sorted in descending order according to* W **do**
  $E \leftarrow E \cup f$
  $q_f \leftarrow p_f$
  if $R(q, u) \geq R(p, u)$ **then**
    break
return E

This approach output contrastive explanations since it answer the question "why p and not q"?. This explanation is also selected as only small set of features values presented. Furthermore, the explanations are expected to have high fidelity, since they do not try to approximate the original models but examine changes in its output.

2.4 Experimental Evaluation

2.4.1 Experimental Design

We evaluated the explanation generated by CX-RS through a survey on Mechanical Turk [22]. This survey requires an item feature-based dataset with user ratings. Unfortunately, many of the available rating datasets focus on content recommendations (e.g., movies and books) and thus they have few numerical features. Therefore, these datasets are less suitable for our setting.

Therefore, we composed our own dataset. To that end, we first collected data on the most popular cell phones in the US in 2022, as mentioned by several leading

websites [1]. The data for each cell phone consists of the most important features such as performance rating (AnTuTu), memory size, camera's resolution, battery size, screen size, release date, etc. We collected the price of each cell phone from Amazon and Best-Buy (in Aug 22). In order to elicit the ratings, we conducted a survey on Mechanical Turk. Each participant was presented with 10 random cell phones, and she was asked to indicate how likely she is to purchase each of the cell phones at the given price, on a scale from 1 (very unlikely) to 10 (very likely). We also asked each participant to add personal information: age, gender and occupation. Overall, in our dataset there are 34 cell phones with 13 features.

In our main survey, we evaluated the explanation generated by CX-RS. We ran a similar survey on Mechanical Turk, in which the participants were asked to rate 10 cell phones and provide demographic information. Then, two cell phones were presented. The first is the most recommended cell phone for the participant, selected by a state-of-the-art recommendation system. The participant was told that the second cell phone was preferred by users with preferences similar to his. We output the recommended cell phone using two state-of-the-art recommendation system models.

- *MeLU* [52], a meta-learning-based recommendation system that designed to alleviate the cold-start problem, which is the problem of recommending items to users who consumed only a few items. From meta-learning (learning to learn), which can rapidly adopt new task with a few examples, MeLU can estimate the user preferences using only a few items.
- *Wide & Deep* [53], which is a recommendation system that combines the benefits of a linear ("wide") model and a deep learning model. A linear model with a wide set of features can memorize the interaction between the features, while a deep neural network can generalize better to unseen feature combinations through low-dimensional dense embeddings learned for the sparse features. Wide & Deep learning jointly trains wide linear models and deep neural networks to combine the benefits of memorization and generalization for recommendation systems.

We compare our explanation with two baselines. The first baseline randomly select features whose values in p and q are not equal. In the second baseline a linear regression model is trained on all the items that a user rated, and the features with the highest absolute weights will be selected. The features are ordered in descending order according to the weights of the linear model. In both baselines, the number of features selected will be the same as the number of features our explanation select to present to the user.

For the first survey, in which we collected ratings on cellphones, we set the reward to \$0.5. We recruited 100 participants, of them, 51 were males, 46 females, and 3 participants who chose not to specify their gender.The average age of the participants was 36. In the main survey, which compared CX-RS with the two other baselines, the reward for each participant was \$0.7. Here we also recruited 100 participants, half received a recommendation according to Melu and the other half

[1] https://www.theverge.com/22163811/best-phone.
https://www.techadvisor.com/article/724318/best-smartphone.html.
https://www.tomsguide.com/best-picks/best-phones.

according to Wide & Deep. Of the 100 participants 47 were males and 53 females. The average age of the participants was 38. In both surveys we set a requirement on Mechanical Turk that the approval rate of the workers must be at least 99% and did not require the Turkers to be masters.

2.4.2 Results

Figures 2.1,2.2 compares the performance, in terms of average ratings, of CX-RS with that of the baselines in both Melu and Wide & Deep recommendation systems. As depicted by the figure, the explanations that were generated by CX-RS outperformed all other explanations. These differences are statistically significant ($p < 0.01$; using a student t-test). That is, the human participants, when presented with explanations according to our method, are more convinced that the recommended item is better for them, than when using contrastive explanations generated by other methods.

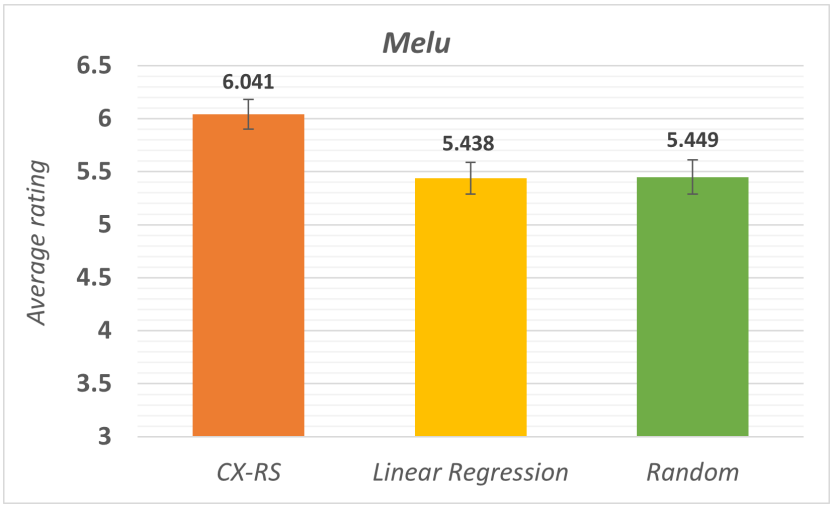

FIGURE 2.1: Average rating for Melu for every type of explanation. Error bars present the standard error.

FIGURE 2.2: Average rating for Wide & Deep for every type of explanation. Error bars present the standard error.

2.5 Future Work

In future work, we intend to allow the user to further ask why the system did not recommend another, different item, q'. A naive approach, which can be used as a baseline, is to use our approach from the first phase and find features that explain the advantages of p over q'. However, we hypothesize that an algorithm that considers also the previously queried item q along with the previously provided explanation, will result in better explanations for the user. Clearly, the interaction between the user and the system may include several contrastive queries, and we intend to develop algorithms that consider *all* previously contrastive queries and previously provided explanations to generate a new explanation.

Another setting that we intend to tackle in the future is to allow the user to post multiple contrastive queries simultaneously. Our algorithm should provide a more holistic explanation, accounting for all the contrastive queried items together.

Chapter 3

Conclusions

In the first part of this thesis, we studied explanations to Shapley-value allocations. 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.

We first define ETX games, which represent common scenarios and their Shapley allocation is easy to understand. We then presented X-SHAP, an algorithm that decomposes any coalitional game into ETX sub-games and generates a brief verbal explanation to every sub-game. We showed 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.

In the second part we presented CX-RS, an algorithm that allows users to pose contrastive questions to recommendation systems. CX-RS can be embedded in any feature-based recommendation system models. This will enable users to pose contrastive questions to a recommendation system, a significant feature missing in current recommendation systems. Our method make the systems more attentive and not only provide a recommendation but also allow the user to interact with them and question their decisions. We showed that when applying our method, humans are more convinced that the recommended item is better for them, than when using contrastive explanations generated by other methods.

Bibliography

[1] John C. Harsanyi. “A Bargaining Model for the Cooperative n-Person Game”. In: *Contributions to the Theory of Games IV*. Ed. by Albert William Tucker and Robert Duncan Luce. Vol. 2. Princeton University Press, 1959, pp. 325–356.

[2] John C. Harsanyi. “A Simplified Bargaining Model for the n Person Cooperative Game”. In: *International Economic Review* 4 (1963), pp. 194–220.

[3] D. Schmeidler. “The Nucleolus of a Characteristic Function Game”. In: *Siam Journal on Applied Mathematics* 17 (1969), pp. 1163–1170.

[4] Lloyd S Shapley. “A value for n-person games”. In: *Contributions to the Theory of Games* 2.28 (1953), pp. 307–317.

[5] Eyal Winter. “The Shapley value”. In: *Handbook of game theory with economic applications* 3 (2002), pp. 2025–2054.

[6] Sergiu Hart. “Shapley value”. In: *Game Theory*. Springer, 1989, pp. 210–216.

[7] Geoffroy de Clippel. “Membership separability: A new axiomatization of the Shapley value”. In: *Games and Economic Behavior* 108 (2018), pp. 125–129.

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

[9] Ariel D Procaccia. “Axioms should explain solutions”. In: *The Future of Economic Design*. Springer, 2019, pp. 195–199.

[10] Jonathan Goldman and Ariel D. Procaccia. “Spliddit: Unleashing Fair Division Algorithms”. In: *SIGecom Exch.* 13.2 (2015), pp. 41–46.

[11] Meir Nizri, Amos Azaria, and Noam Hazon. “Improving the Perception of Fairness in Shapley-Based Allocations”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 44. 44. 2022.

[12] Mark G Core et al. “Building explainable artificial intelligence systems”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2006, pp. 1766–1773.

[13] David Gunning et al. “XAI—Explainable artificial intelligence”. In: *Science Robotics* 4.37 (2019).

[14] Amina Adadi and Mohammed Berrada. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 6 (2018), pp. 52138–52160.

[15] Derek Doran, Sarah Schulz, and Tarek R Besold. “What does explainable AI really mean? A new conceptualization of perspectives”. In: *arXiv preprint arXiv:1710.00794* (2017).

[16] Andreas Holzinger et al. “What do we need to build explainable AI systems for the medical domain?” In: *arXiv preprint arXiv:1712.09923* (2017).

[17] Sarit Kraus et al. “AI for explaining decisions in multi-agent environments”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 13534–13538.

[18] Olivier Cailloux and Ulrich Endriss. “Arguing about Voting Rules”. In: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2016, pp. 287–295.

[19] Dominik Peters et al. “Explainable Voting”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 1525–1534.

[20] Ari Stern and Alexander Tettenhorst. “Hodge decomposition and the Shapley value of a cooperative game”. In: *Games and Economic Behavior* 113 (2019), pp. 186–198.

[21] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *NeurIPS*. Vol. 31. 2017, pp. 4768–4777.

[22] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. “Running experiments on amazon mechanical turk”. In: *Judgment and Decision making* 5.5 (2010), pp. 411–419.

[23] Ankur Joshi et al. “Likert scale: Explored and explained”. In: *British Journal of Applied Science & Technology* 7.4 (2015), p. 396.

[24] Yazan Mualla et al. “The Quest of Parsimonious XAI: a Human-Agent Architecture for Explanation Formulation”. In: *Artificial Intelligence* 302 (2021), pp. 1–26.

[25] Frank E Harrell. “Ordinal logistic regression”. In: *Regression modeling strategies*. Springer, 2015, pp. 311–325.

[26] Chayaporn Suphavilai, Denis Bertrand, and Niranjan Nagarajan. “Predicting Cancer Drug Response using a Recommender System”. In: *Bioinformatics* 34 (2018), pp. 3907–3914.

[27] Dávid Zibriczky. “Recommender Systems meet Finance: a Literature Review”. In: *FINREC*. 2016.

[28] Nava Tintarev and Judith Masthoff. “Explaining Recommendations: Design and Evaluation”. In: *Recommender Systems Handbook*. Springer, 2015, pp. 353–382.

[29] Behnoush Abdollahi and Olfa Nasraoui. “Using Explainability for Constrained Matrix Factorization”. In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 2017, pp. 79–83.

[30] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. “Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 717–725.

[31] Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretable Machine Learning”. In: *arXiv: Machine Learning* (2017).

[32] Robert Nozick. *Philosophical explanations*. Harvard University Press, 1983.

[33] Tania Lombrozo. "The structure and function of explanations". In: *Trends in Cognitive Sciences* 10 (2006), pp. 464–470.

[34] Peter Lipton. "Contrastive Explanation". In: *Royal Institute of Philosophy Supplement* 27 (1990), pp. 247–266.

[35] Zachary C. Lipton. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[36] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. "Explaining collaborative filtering recommendations". In: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 2000, pp. 241–250.

[37] Bruce Ferwerda, Kevin Swelsen, and Emily Yang. "Explaining content-based recommendations". In: *New York* (2018), pp. 1–24.

[38] Lara Quijano-Sanchez et al. "Make it personal: a social explanation system applied to group recommendations". In: *Expert Systems with Applications* 76 (2017), pp. 36–48.

[39] Lei Li, Li Chen, and Ruihai Dong. "CAESAR: context-aware explanation based on supervised attention for service recommendations". In: *Journal of Intelligent Information Systems* 57 (2021), pp. 147–170.

[40] Georgina Peake and Jun Wang. "Explanation mining: Post hoc interpretability of latent factor models for recommendation systems". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 2060–2069.

[41] Caio Nóbrega and Leandro Marinho. "Towards explaining recommendations through local surrogate models". In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. 2019, pp. 1671–1678.

[42] Alexandre Chanson, Nicolas Labroche, and Willème Verdeaux. "Towards local post-hoc recommender systems explanations". In: *Proceedings of the 23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP)*. 2021.

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[44] Mingming Guo et al. "Online Product Feature Recommendations with Interpretable Machine Learning". In: *arXiv preprint arXiv:2105.00867* (2021).

[45] Dorin Shmaryahu, Guy Shani, and Bracha Shapira. "Post-hoc Explanations for Complex Model Recommendations using Simple Methods." In: *Proceedings of the 7th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*. 2020, pp. 26–36.

[46] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harvard journal of law & technology* 31 (2017), pp. 841–887.

[47] Vassilis Kaffes, Dimitris Sacharidis, and Giorgos Giannopoulos. “Model-agnostic counterfactual explanations of recommendations”. In: *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 2021, pp. 280–285.

[48] Azin Ghazimatin et al. “PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems”. In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. 2020, pp. 196–204.

[49] Jinfeng Zhong and Elsa Negre. “Shap-Enhanced Counterfactual Explanations for Recommendations”. In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. Association for Computing Machinery, 2022, pp. 1365–1372.

[50] Chen He, Denis Parra, and Katrien Verbert. “Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities”. In: *Expert Systems with Applications* 56 (2016), pp. 9–27.

[51] Chongming Gao et al. “Advances and Challenges in Conversational Recommender Systems: A Survey”. In: *AI Open* 2 (2021), pp. 100–126.

[52] Hoyeop Lee et al. “MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019).

[53] Heng-Tze Cheng et al. “Wide & Deep Learning for Recommender Systems”. In: *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. 2016, pp. 7–10.

אוניברסיטת אריאל בשומרון

תקציר

הפקולטה למדעי הטבע

המחלקה למדעי המחשב

מאסטר

הסברה של הקצאה הוגנת והמלצות

על ידי מאיר נזרי

ככל שבינה המלאכותית (AI) הופכת מתקדמת ומשמעותית יותר בהיבטים נרחבים של חיי היומיום שלנו, בני אדם מאותגרים לעקוב ולהבין את האופן שבו האלגוריתם הגיע לתוצאה. ישנם יתרונות רבים להבנה כיצד מערכת תומכת AI הובילה לתוצאה ספציפית. הסבר איכותי לתוצאה מאפשר לאנשים המושפעים מהחלטה של המערכת לערער או לשנות את התוצאה הזו, ייתכן שיהיה צורך גם לעמוד בתקנים רגולטוריים, או עזרה למפתחים להבטיח שהמערכת פועלת כמצופה. בתזה זו פיתחנו שיטות להסבר אלגוריתמי בינה מלאכותית בשני תחומים משותפים: תורת המשחקים ומערכות המלצה.

בפרק הראשון של התזה, אנו מסבירים הקצאות לפי ערך שאפלי. ערך-שאפלי הוא אחת משיטות החלוקה הנורמטיביות החשובות ביותר בתורת המשחקים השיתופיים, המקיימת אקסיומות בסיסיות. עם זאת, הקצאות מסוימות לפי ערך ערך-שאפלי עשויות להיראות לא הוגנות לבני אדם. בתזה זו, אנו פיתחנו שיטה אוטומטית המייצרת הסברים אינטואיטיביים להקצאה מבוססת ערך-שאפלי, המשתמשת באקסיומות הבסיסיות. בהינתן כל משחק קואליציוני, השיטה שלנו מפרקת אותו לתתי משחקים, שעבורם קל לייצר הסברים מילוליים, ומראה שהמשחק הנתון מורכב מתתי המשחקים. מאחר שההקצאה עבור כל תת-משחק נתפסת כהוגנת, ההקצאה מבוססת ערך-שאפלי עבור המשחק הנתון אמורה גם כן להיראות הוגנת.

בפרק הראשון של התזה, אנו מסבירים את פלט של מערכות ההמלצה. מערכות המלצות נמצאות בשימוש נרחב והן קיימות ביישומים רבים, כגון המלצת סרטים, מכירת מוצרים וספקי תוכן. עם זאת, מערכות ההמלצה הנוכחיות הן בדרך כלל חסרות את היכולת להסביר את המלצותיהן או לאפשר למשתמשים לחקור אותן. אנו פיתחנו שיטה אוטומטית המייצרת הסברים מנוגדים למערכות המלצות המבוססות על תכונות הפריטים והעדפות המשתמשים. כלומר, לאחר קבלת המלצה (למשל, לקנות פלאפון Samsung S22), תהיה למשתמשים אפשרות לשאול את המערכת מדוע היא לא המליצה על פריט אחר ספציפי (למשל, Xiaomi 12). באמצעות השיטה שלנו למערכת ההמלצה תהיה היכולת להשיב עם הסבר אישי, משמעותי ומשכנע (למשל, נראה שמצלמה טובה מאוד חשובה למשתמש ספציפי, וה-Samsung S22 כולל מצלמה טובה יותר מה-Xiaomi 12).

בשני המחקרים, ערכנו ניסויים עם משתתפים אנושיים והראינו שכאשר מיישמים את השיטות שלנו, בני אדם משוכנעים יותר שהפלט של ה-AI טובה יותר עבורם מאשר כאשר לא משתמשים באף הסבר או כאשר משתמשים בהסברים שנוצרו על ידי שיטות אחרות.

אוניברסיטת אריאל בשומרון

עבודת תזה

הסברה של הקצאה הוגנת והמלצות

מחבר:

מאיר נזרי

מנחים:

פרופ' עמוס עזריה

ד"ר נועם חזון

המחלקה למדעי המחשב

31 באוקטובר 2022

ה' חשוון תשפ"ג

אוניברסיטת אריאל בשומרון

עבודת תזה

הסברה של הקצאה הוגנת והמלצות

מחבר:

מאיר נזרי

מנחים:

פרופ' עמוס עזריה

ד"ר נועם חזון

המחלקה למדעי המחשב

31 באוקטובר 2022

ה' חשוון תשפ"ג