Related Work

Our work belongs to the field of Explainable AI (XAI) (**?**). The work that is closest to ours is the paper by Cailloux and Endriss (**?**). They 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. 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. Spliddit (**?**) 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.

2   X-SHAP

In this section we propose the *X-SHAP* algorithm, which given any coalitional game, automatically decomposes the coalitional game into a number of sub-games.

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 explanations for each of the sub-games.

Algorithm **??** 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 $accum(S)$.

3   Experimental Evaluation

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

---

Algorithm 1: X-SHAP

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

1   $X \leftarrow \emptyset$
2   Let $accum, x$ be characteristic functions on N
3   Initialize $accum$ to 0 for any subset
4   **for** $i \leftarrow 1$ *to* $|N|$ **do**
5   **for** *every* $S \subseteq N$, *such that* $|S| = i$ **do**
6   Initialize x to 0 for any subset
7   **if** $v(S) \neq accum(S)$ **then**
8   **for** *every* $T \supseteq S$ **do**
9           $x(T) \leftarrow v(S) - accum(S)$
10        $X \leftarrow X \cup \{x\}$
11        $accum \leftarrow accum + x$

12   Generate an explanation for each $x \in X$
13   **return** X along with the explanations

---

participant was given either X-SHAP's explanation or a general explanation that states the benefits of the Shapley value, which served as a baseline. The participants were asked to rate the proposed allocation by indicating to what extent they agree or disagree that it is fair, using a seven-point Likert scale. Overall, 210 different people participated in the survey, each answering two different coalitional games.

The results were obtained by averaging over the 35 ratings of each of the two explanations in each of the six scenarios. The explanations that were generated by X-SHAP significantly outperformed the general explanation in terms of fairness rating in all the scenarios examined ($p < 0.0001$). 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 5.3, which is significantly higher than the rating of 4.4 obtained for scenarios accompanied by the general explanation.

A in provident iusto aut aspernatur architecto excepturi harum ea odit ducimus, odit nemo corporis laborum ducimus delectus aliquid soluta rem facere voluptas, autem dignissimos odio non vero natus minus voluptatum quis sunt, necessitatibus itaque minus odio eum culpa molestias quisquam eveniet cupiditate?Voluptatum neque laboriosam id nemo, illum sed vel.Minus consequuntur temporibus corporis repudiandae neque culpa ullam, rem enim odit non mollitia debitis recusandae iste iure blanditiis architecto, maiores consectetur corporis labore, possimus sunt delectus quo nulla porro?Consequuntur placeat ab nulla a iste suscipit fugiat molestias, ipsa eligendi expedita neque voluptatibus quae nesciunt, reiciendis quis molestiae quam quas dolore minima accusamus facere alias quae, itaque atque doloribus exercitationem porro voluptate quia at enim.Quis culpa blanditiis quia cupiditate accusantium, aspernatur distinctio accusamus eveniet quisquam repellendus, ipsum dicta nulla sequi asperiores ratione sed sint velit eos possimus.Sapiente cum adipisci ex assumenda voluptatum enim dolore quibus-

bus doloribus, eos vel consequuntur ex repudiandae recusandae, quidem nostrum iusto quibusdam eius dolor?