Testing Axioms Against Human Reward Divisions in Cooperative Games

Greg d'Eon
University of British Columbia
gregdeon@cs.ubc.ca

Kate Larson
University of Waterloo
kate.larson@uwaterloo.ca

ABSTRACT

Axiomatic approaches are an appealing method for designing fair algorithms, as they provide formal structure for reasoning about and rationalizing individual decisions. However, to make these algorithms useful in practice, their axioms must appropriately capture social norms. We explore this tension between fairness axioms and socially acceptable decisions in the context of cooperative game theory. We use two crowdsourced experiments to study people's impartial reward divisions in cooperative games, focusing on games that systematically vary the values of the single-player coalitions. Our results show that people select rewards that are remarkably consistent, but place much more emphasis on the single-player coalitions than the Shapley value does. Further, their reward divisions violate both the null player and additivity axioms, but support weaker axioms. We argue for a more general methodology of testing axioms against experimental data, retaining some of the conceptual simplicity of the axiomatic approach while still using people's opinions to drive the design of fair algorithms.

KEYWORDS

Cooperative games; Shapley value; human behaviour

ACM Reference Format:
Greg d'Eon and Kate Larson. 2020. Testing Axioms Against Human Reward Divisions in Cooperative Games. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

As we give algorithms the power to make high-stakes decisions, it is important to ensure that these decisions are fair. There are several ways to design such an algorithm. One is the axiomatic approach, where the algorithm makes provably fair decisions that satisfy a number of mathematical axioms. These axioms are attractive, leading to conceptually simple algorithms and, in some cases, the ability to explain individual outcomes [11, 34]. However, when they fail to capture social norms, they often produce decisions that are deemed to be unfair [23, 24]. When this approach fails, the main alternative is to set aside these theoretical guarantees, elicit stakeholders' opinions, and encode these into the model's behaviour. This idea drives modern AI techniques for ethical and fair decision making [10, 21, 26, 31].

This conflict between fairness axioms and socially acceptable decisions is present in cooperative game theory. Here, the most celebrated reward division method is the Shapley value [36], which is the unique division that satisfies four fairness axioms of efficiency, symmetry, null players, and additivity. However, there is little reason to believe that these axioms align with human fairness principles. Nowak and Radzik [32] suggested that the Shapley value is only suitable for pure economic situations, and different axioms might be necessary to represent important social or psychological aspects. This criticism inspired alternative reward divisions, such as the solidarity value [33] and the egalitarian Shapley values [18], that capture these more "human" qualities.

Unfortunately, there is little empirical data on cooperative games to validate whether these alternative fair divisions truly are more "human". The most relevant experimental work is De Clippel and Rozen [8], who studied how impartial decision makers divided rewards in a limited set of games. Their results suggested that people selected convex combinations of equal divisions and the Shapley values. This trend implies that most people followed symmetry, efficiency, and additivity, only breaking the null player axiom.

In this paper, we apply De Clippel and Rozen's approach to a broader set of games. Specifically, we design two sets of 3-player cooperative games that systematically vary the values of the single-player coalitions, and we use two crowdsourced experiments (n = 75 and 74) to find how people select impartial reward divisions in these games. Our results show that people select rewards that are remarkably consistent, but have little to do with the Shapley value: they place much more emphasis on the single-player coalitions. Further, their reward divisions violate both the null player and additivity axioms, but support weaker monotonicity axioms.

We make two main contributions. First, our data provides new insights about the gap between people's opinions and fairness axioms from cooperative game theory. Second, we propose the use of our methodology in a broader set of social choice problems to identify weaker axioms that are consistent with experimental data. We argue that these weaker axioms are useful for representing invariants in people's opinions, even if they are not used to select a single provably fair decision. In this way, we can retain some of the conceptual simplicity of the axiomatic approach while still using people's opinions to drive the design of fair algorithms.

1.1 Related Work

We situate our work in the literature on human behaviour in cooperative games. The earliest experimental work is by Kalisch et al. [22], who studied face-to-face bargaining with 4 to 7 people. Though their main focus was on the bargaining dynamics, they identified several human factors: players often split their rewards equally, and powerful players rarely took advantage of their position.

Most of the experimental work following Kalisch et al. is characterized by two features. First, it focuses on bargaining, with participants discussing coalitions and reward divisions while acting

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

as players in the games. Second, it uses *zero-normalized* games, where players must form coalitions to earn rewards. One notable exception uses non-zero-normalized games [19], but still focuses on bargaining in these games. For comprehensive surveys of this work, we refer readers to Kahan and Rapoport [20] and Maschler [29]. Some cooperative game theoretic predictions have also been validated outside of laboratory studies [38]. Recent work has continued to focus on bargaining, using structured protocols [4, 30] or computer agents [41].

De Clippel and Rozen's experiment [8] is the most relevant to our work. In their experiment, three "recipients" earned baskets of items by answering trivia questions. Then, impartial "decision makers" divided monetary rewards between the recipients based on the baskets' values. They concluded that humans select convex combinations of an equal split and the Shapley value. To our knowledge, their work is the first where the participants dividing the rewards are impartial to the divisions. However, the games were zero-normalized, as the recipients' baskets were worthless alone.

We note some parallels between our research and other empirical work. First, behavioural economics often uses experiments to identify gaps between theory and human behaviour. Halevy's experiment on subjective uncertainty [13] is a classic example. Computational models, such as those in behavioural game theory [5, 14, 39], try to understand this gap by predicting people's strategic decisions. Second, in fair division problems, human perceptions of fairness are rarely aligned with theoretical properties like envy freeness. Instead, they are often influenced by interpersonal comparisons [16], reference points such as equal divisions [15], multiple concepts of fairness [23], and understanding of the fair division system [25]. Finally, empirical work in machine learning has compared clustering quality metrics against human evaluations [27]. We take inspiration from these approaches, but to our knowledge, none of these models can directly applied to cooperative games.

2 VALUES FOR COOPERATIVE GAMES

We begin by describing cooperative game theory concepts that we use to motivate our experiments. A *transferable utility game* $G = (N, f)$ consists of a set of n players, N, and a characteristic function $f \in \mathbb{R}^{2^N}$, assigning a reward $f(C)$ to each coalition $C \subseteq N$. We typically require $f(\emptyset) = 0$: a coalition with no players earns no reward. In this paper, we restrict our attention to 3-player transferable utility games, where $N = \{A, B, C\}$, and we refer to the characteristic function f as a "game". For convenience, we often write the set $\{i\}$ as i and $\{i, j\}$ as ij.

A player i's *marginal contribution* to a coalition $C \subseteq N \setminus i$ is $mc(i, f, C) = f(C \cup i) - f(C)$: the amount of value that a player adds to a coalition by joining it. Players i and j are *symmetric* if $mc(i, f, C) = mc(j, f, C)$ for all $C \subseteq N \setminus ij$, and player i is a *null player* if $mc(i, f, C) = 0$ for all $C \subseteq N \setminus i$. A game is *monotonic* if every marginal contribution is non-negative, and a game is *zero-normalized* if $f(i) = 0$ for all players $i \in N$.

A *value* is a function $v : \mathbb{R}^{2^N} \to \mathbb{R}^N$ that assigns a reward $v_i(f)$ to each of the players i in a game f. A simple example is the equal division value ED, which gives each player an equal fraction of the total: $ED_i(f) = f(N)/n$. The most celebrated value is the Shapley value [36], which is the unique value Sh that satisfies four axioms:

- *Efficiency*: $\sum_i Sh_i(f) = f(N)$.
- *Symmetry*: if players i and j are symmetric, $Sh_i(f) = Sh_j(f)$.
- *Null players*: if player i is a null player, $Sh_i(f) = 0$.
- *Additivity*: for any two games f and g, let $(f + g)(C) = f(C) + g(C)$ for all $C \subseteq N$. Then, $Sh(f + g) = Sh(f) + Sh(g)$.

This value has a simple interpretation: considering all possible orders that the players could form the grand coalition, each player is given their average marginal contribution. Equivalently,

$$Sh_i(f) = \sum_{C \subseteq N \setminus i} \frac{|C|!(n - |C| - 1)!}{n!} mc(i, f, C).$$

A number of modifications to the Shapley value have been proposed. Each of them satisfy efficiency, symmetry, and additivity, modifying the null player axiom. One is the family of *egalitarian Shapley values* [6, 18], which are convex combinations of the equal division and Shapley values:

$$Sh^\alpha(f) = \alpha Sh(f) + (1 - \alpha) ED(f).$$

Here, the parameter α describes a social norm of equality: $\alpha = 0$ gives the equal division, while $\alpha = 1$ recovers the Shapley value. Another is Nowak and Radzik's *solidarity value* [32], which is

$$Sol_i(f) = \sum_{C \ni i} \frac{(n - |C|)!(|C| - 1)!}{n!} A^f(C)$$

where $A^f(C) = \frac{1}{|C|} \sum_{i \in C} mc(i, f, C)$ is the average marginal contribution of any player to C. Nowak and Radzik suggest that the solidarity value is more human, capturing some subjective psychological aspects of the game, while the Shapley value is the "pure economic" solution. Finally, these two ideas are generalized by the family of *procedural values* [28, 35], which are

$$P^s(f) = \sum_{C \subseteq N \setminus i} \frac{|C|!(n - |C| - 1)!}{n!} \left[s_{|C|+1} f(C \cup i) - s_{|C|} f(C) \right].$$

Procedural values are described by tuples $s = (s_1, \ldots, s_{n-1})$, with the convention $s_0 = s_n = 1$. Each term s_k is a measure of equality: when a player joins a coalition of size k, they keep a fraction s_k of their marginal contribution, splitting the remaining fraction $(1 - s_k)$ equally among the other players. The family of procedural values includes all of the values described previously: $ED(f)$ has $s_k = 0$, $Sh(f)$ has $s_k = 1$, $Sh^\alpha(f)$ has $s_k = \alpha$, and $Sol(f)$ has $s_k = \frac{1}{k+1}$.

We note that there is a separate set of solution concepts that focus on *stability* of the players' rewards, such as the core, the kernel, and the bargaining set [7]. These concepts are based on the behaviour of rational players instead of fair and impartial reward divisions, which is the focus of this paper.

3 METHOD

We used two crowdsourced experiments to study people's reward divisions in 3-player cooperative games. In the first experiment, we tested whether the values of the 1- or 2- player coalitions had more influence on people's reward divisions. We took this idea further in the second experiment, testing whether the 1-player coalitions could "mislead" people's reward divisions.

3.1 Games

We designed the games for our experiments to vary two main factors. First, we considered different distributions of power between the players, making games where some players were more valuable than others. Second, we realized these differences in a variety of ways by carefully controlling each player's marginal contributions.

Experiment 1: We began by designing three games with a Shapley value of [25, 25, 10]. Since player C is less valuable than players A and B, we call this reward division the 1-Worse value. In these three games, we controlled the players' marginal contributions through the values of the Solo coalitions, the Pair coalitions, or Both. In the Solo game, we chose identical values for each of the 2-player coalitions, so the players only differed in the values of their 1-player coalitions. Conversely, in the Pair game, the 1-player coalitions had identical values, and the only differences were in the 2-player coalition values. In the Both game, these two coalition sizes had identical effects on the Shapley values. We repeated this process for Shapley values of [30, 15, 15] (1-Better) and [30, 20, 10] (Distinct). Finally, we included a purely additive game with $Sh = [30, 20, 10]$ and a purely symmetric game. These 11 games are listed in Table 1.

Experiment 2: We considered the 1-Worse value again, but we chose six games with 1-player coalition values that made player A appear more valuable than player B. For the first three of these games, we gave player A's solo coalition a value of 2, 5, or 10, fixing player B and C's solo values to 0; we call these the Zeros2, Zeros5, and Zeros10 games. For the latter three, we made player A's solo value 15 higher than players B and C, choosing solo values that summed to 30, 45, or 60; we call these the Sum30, Sum45, and Sum60 games. We chose six analogous games for the 1-Better value, choosing 1-player coalition values that made players A and B appear equal.

We also designed four games with a Shapley value of [40, 20, 0]. Since player C is a null player, we call this value the 1-Null value. We chose these games by setting player B's solo coalition value to 0 (Zeros) or having player A and B's solo values sum to 40, 50, or 60 (Sum40, Sum50, Sum60). Finally, we included the Symmetric game again. All 17 games are listed in Table 2.

	Characteristic function (f)						$Sh(f)$		
Game	A	B	C	AB	AC	BC	A	B	C
1-Worse-Solo	40	40	10	60	60	60	25	25	10
1-Worse-Both	15	15	0	45	30	30			
1-Worse-Pair	0	0	0	45	15	15			
1-Better-Solo	40	10	10	60	60	60	30	15	15
1-Better-Both	15	0	0	45	45	30			
1-Better-Pair	0	0	0	45	45	15			
Distinct-Solo	40	20	0	60	60	60	30	20	10
Distinct-Both	20	10	0	60	50	40			
Distinct-Pair	0	0	0	60	40	20			
Additive	30	20	10	50	40	30	30	20	10
Symmetric	20	20	20	40	40	40	20	20	20

Table 1: The 11 games used in experiment 1. All games have $f(\emptyset) = 0$ and $f(\mathbf{ABC}) = 60$.

	Characteristic function (f)						$Sh(f)$		
Game	A	B	C	AB	AC	BC	A	B	C
1-Worse-Zeros2	2	0	0	40	10	12	25	25	10
1-Worse-Zeros5	5	0	0	40	10	15			
1-Worse-Zeros10	10	0	0	40	10	20			
1-Worse-Sum30	20	5	5	60	30	45			
1-Worse-Sum45	25	10	10	60	30	45			
1-Worse-Sum60	30	15	15	60	30	45			
1-Better-Zeros2	2	2	0	38	40	10	30	15	15
1-Better-Zeros5	5	5	0	35	40	10			
1-Better-Zeros10	10	10	0	30	40	10			
1-Better-Sum30	15	15	0	45	60	30			
1-Better-Sum45	20	20	5	45	60	30			
1-Better-Sum60	25	25	10	45	60	30			
1-Null-Zeros	20	0	0	60	20	0	40	20	0
1-Null-Sum40	30	10	0	60	30	10			
1-Null-Sum50	35	15	0	60	35	15			
1-Null-Sum60	40	20	0	60	40	20			
Symmetric	20	20	20	40	40	40	20	20	20

Table 2: The 17 games used in experiment 2. All games have $f(\emptyset) = 0$ and $f(\mathbf{ABC}) = 60$.

3.2 Experiments

Participants: We hired participants from Mechanical Turk. For experiment 1, we posted human intelligence tasks (HITs) titled "Divide rewards in fictional scenarios (10 mins)" with a payment of $1.25 USD. For experiment 2, we changed these values to 15 minutes and $1.75 USD. We used Mechanical Turk's qualification system to restrict the HIT to workers who were located in the United States, had at least 1000 approved HITs with 95% or higher approval rate, and had not previously started the experiment.

Task: During the experiment, participants were presented with a series of scenarios about three fictional characters – Alice, Bob, and Charlie – playing a video game online. Each of these scenarios was associated with one of the cooperative games. We displayed the details of the game in a colour-coded table, listing the amount of gold that each combination of players could earn. Then, we told workers that all three players chose to work together, and we asked how the gold should be divided. Workers entered their responses by adjusting three sliders and clicking the submit button. The interface disabled the submit button as long as there was a surplus, only allowing efficient responses to be submitted. The experiment interface is shown in Figure 1.

Procedure: After workers accepted the HIT, they filled out a consent form and completed a brief tutorial. In this tutorial, we described the interface and asked comprehension questions about the reward displays. Then, workers completed several rounds of the task, with each round corresponding to one of the games above. We randomized the order of the games. We also randomly labelled players A, B, and C as Alice, Bob, and Charlie in each game. Finally, workers received a confirmation code and submitted the HIT.

Figure 1: The task interface. Participants were presented with a tabular representation of the game and asked to divide the reward between the three players. The "submit" button was enabled when the entire reward was allocated.

4 RESULTS

In both experiments, a total of 100 workers completed the HIT. To deal with workers that submitted low-quality answers (such as [30, 30, 0] in the Symmetric games), we filtered out workers that spent less than 5 seconds on any game, removing 21 workers from each experiment. We also manually removed workers that submitted multiple nonsensical answers (for example, [1, 1, 58] in Distinct-Both). After filtering, we were left with 75 workers in Experiment 1 and 74 workers in Experiment 2. We confirmed that this criteria was appropriate by checking the Symmetric games: the most extreme reward remaining was [20, 22, 18].

4.1 Experiment 1

The rewards that each participant submitted for each game in Experiment 1 are shown in Figure 2. Each of these plots shows the distribution of selected rewards, along with the equal division (red) and the Shapley value (blue).

The majority of the rewards are close to a few key points. The most common is the equal division, which was picked by at least 25 of the 75 participants in each game. Other frequent points are the Shapley value and rewards that are half or double the distance from the equal division to the Shapley value. However, the frequencies of these unequal rewards differ between the games. For all three Shapley values, the Solo games have the most extreme rewards, while the Both and Pair divisions are generally more equal. For instance, in 1-Better-Solo, 14 participants submitted rewards close to [40, 10, 10]; in 1-Better-Both, only 3 such rewards remained.

We confirmed this trend using non-parametric statistical tests. For each division, we calculated the ℓ_1 distance to the equal division, and we compared these distances using Holm-Bonferroni-corrected Wilcoxon signed-rank tests. For all three Shapley values, we found a significant difference between the Solo and Pair games ($p < 0.001$) and between the the Both and Pair games ($p < 0.01$). We also found a significant difference between the 1-Better-Solo and 1-Better-Both games ($p < 0.001$). These results confirm that people gave more equal divisions in the Pair games and more unequal rewards in the Solo games, suggesting that the values of the 1-player coalitions are more important in their reward divisions.

4.2 Experiment 2

The rewards that participants submitted in Experiment 2 are plotted in Figure 3. These rewards show striking differences from the data in Experiment 1. First, few of the rewards lie between the equal division and the Shapley values. Further, in the 1-Worse and 1-Better games, the Shapley values are quite uncommon: in fact, no participants chose the Shapley values in four 1-Better games. However, there is still a clear linear pattern to the rewards in most games. In the 1-Worse games, most of the rewards lie between the equal division and the value [60, 0, 0]; in the 1-Better games, they lie between the equal division and [30, 30, 0]. The 1-Null-Zeros game appears to be similar to the 1-Worse games, with many of the responses giving a disproportionately high amount of reward to player A. Lastly, the other 1-Null games have more rewards close to the Shapley values.

We described these trends using principal component analysis (PCA). For each game, the first principal component of the rewards is plotted as a green lines in Figures 3. Due to the high number of participants selecting equal splits in all games, we plotted these components as passing through the equal division. These components are highly consistent, with nearly identical directions in each 1-Worse game and in each 1-Better game. They also show the differences between the 1-Null games, where the components steadily shift from the extreme value in 1-Null-Zeros towards the set of egalitarian Shapley values in 1-Null-Sum50 and 1-Null-Sum60.

To formally compare the data to the Shapley values, we found bootstrapped 99% confidence intervals for the angles of each of these PCA components. Specifically, to compute one bootstrapped estimate, we sampled 74 points with replacement from our Experiment 2 dataset. We calculated the main PCA component of this resampled data and found the angle of this component as it would appear on a ternary plot. We simulated 10000 bootstrapped samples for each game to get a distribution of the PCA angles, and we took the middle 99% of these angles as the confidence interval. Only three of these confidence intervals, for the 1-Null-SumX games, contain the angle between the equal division and the Shapley value. In contrast to De Clippel and Rozen's findings, our data strongly suggests that egalitarian Shapley values are not a good model for people's reward divisions.

5 DISCUSSION

Our results suggest that people place a great deal of importance on the values of the single-player coalitions in cooperative games, resulting in reward divisions that are substantially different from the Shapley value. In this section, we provide additional insights about these reward divisions, comparing them to alternative solution concepts and fairness axioms.

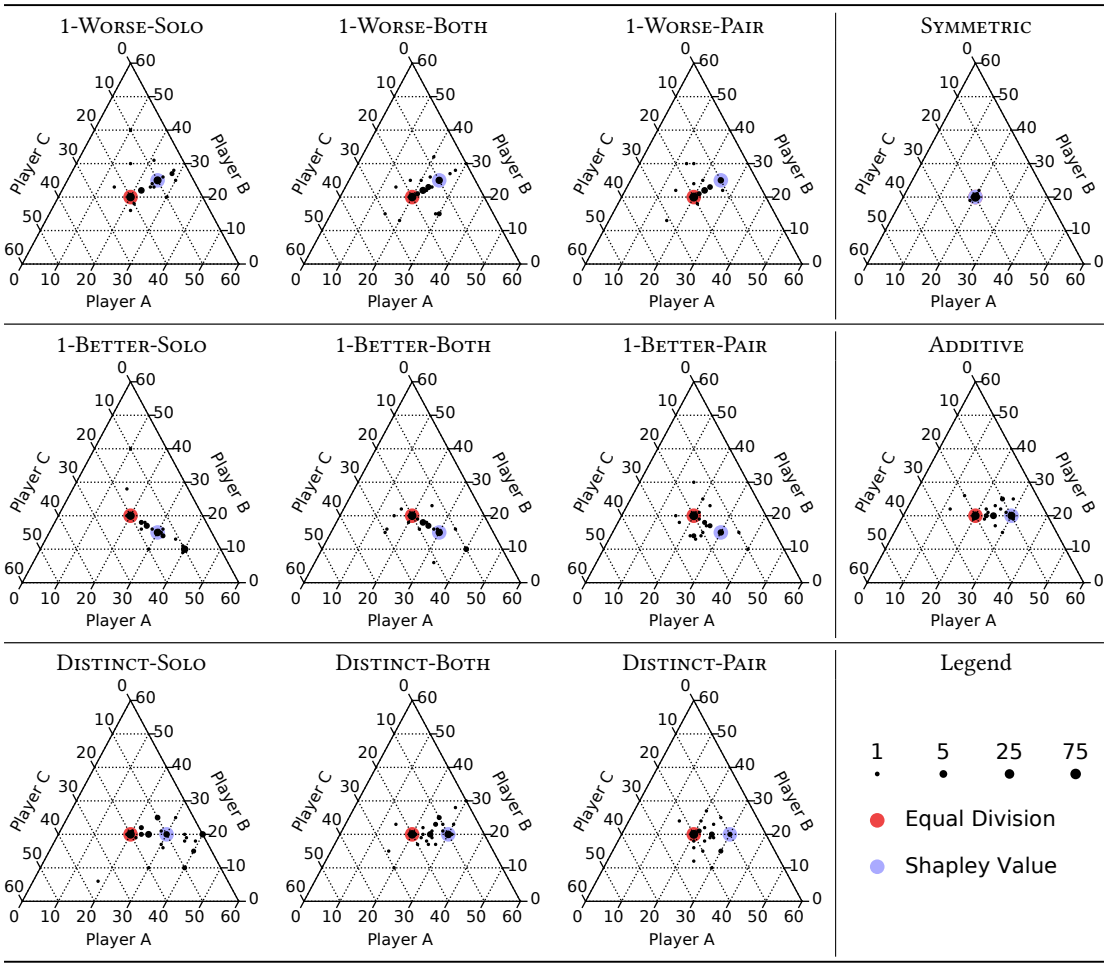

Figure 2: The rewards that participants submitted for each game in Experiment 1. On each plot, $ED(f)$ is circled in dark red, and $Sh(f)$ is circled in light blue.

5.1 Fitting Procedural Values

In De Clippel and Rozen's experiments [8], most people chose egalitarian Shapley values: convex combinations of the Shapley value and the equal division. The natural generalization of egalitarian Shapley values to our data is the family of procedural values, which can treat the 1- and 2-player coalitions differently. Can these procedural values accurately describe our participants' reward divisions?

To answer this question, we tried finding procedural values that accurately described each participant's rewards. We say that the *error* of a pair of parameters (s_1, s_2), relative to a participant's reward divisions, is the maximum ℓ_1 distance between the participant's rewards and the procedural values $P^{(s_1,s_2)}$ across all of the games they saw in the experiment. For each participant, we performed a grid search for the combination of parameters (s_1, s_2) that minimized this error, testing each value in $[-2, 2]$ in steps of 0.05.

We had little success fitting these procedural values to participants' reward divisions. Figure 4 shows the empirical CDF of the errors of these fits. Even allowing for a large ℓ_1 distance of 10 gold pieces, only 40% of the participants' choices could be described by procedural values. Further, most of these participants (25/75 in experiment 1; 19/74 in experiment 2) were the ones who submitted equal divisions in every round of the experiment. We conclude that procedural values are generally not suitable for describing individual people's reward divisions.

5.2 Testing the Shapley Value Axioms

While this negative result shows that people's reward divisions differ from the Shapley value, it does not explain why this difference arises. To gain further insight about this difference, we took inspiration from De Clippel and Rozen's analysis, using our data to check whether people obeyed each of the Shapley value's axioms. In both of our experiments, we only allowed participants to submit efficient rewards, but we made no restrictions related to the other three axioms. Did participants obey these axioms?

Symmetry: Six of the games in Experiment 1 have two symmetric players. In the three 1-Worse games, players A and B are symmetric; in the 1-Better games, players B and C are symmetric.

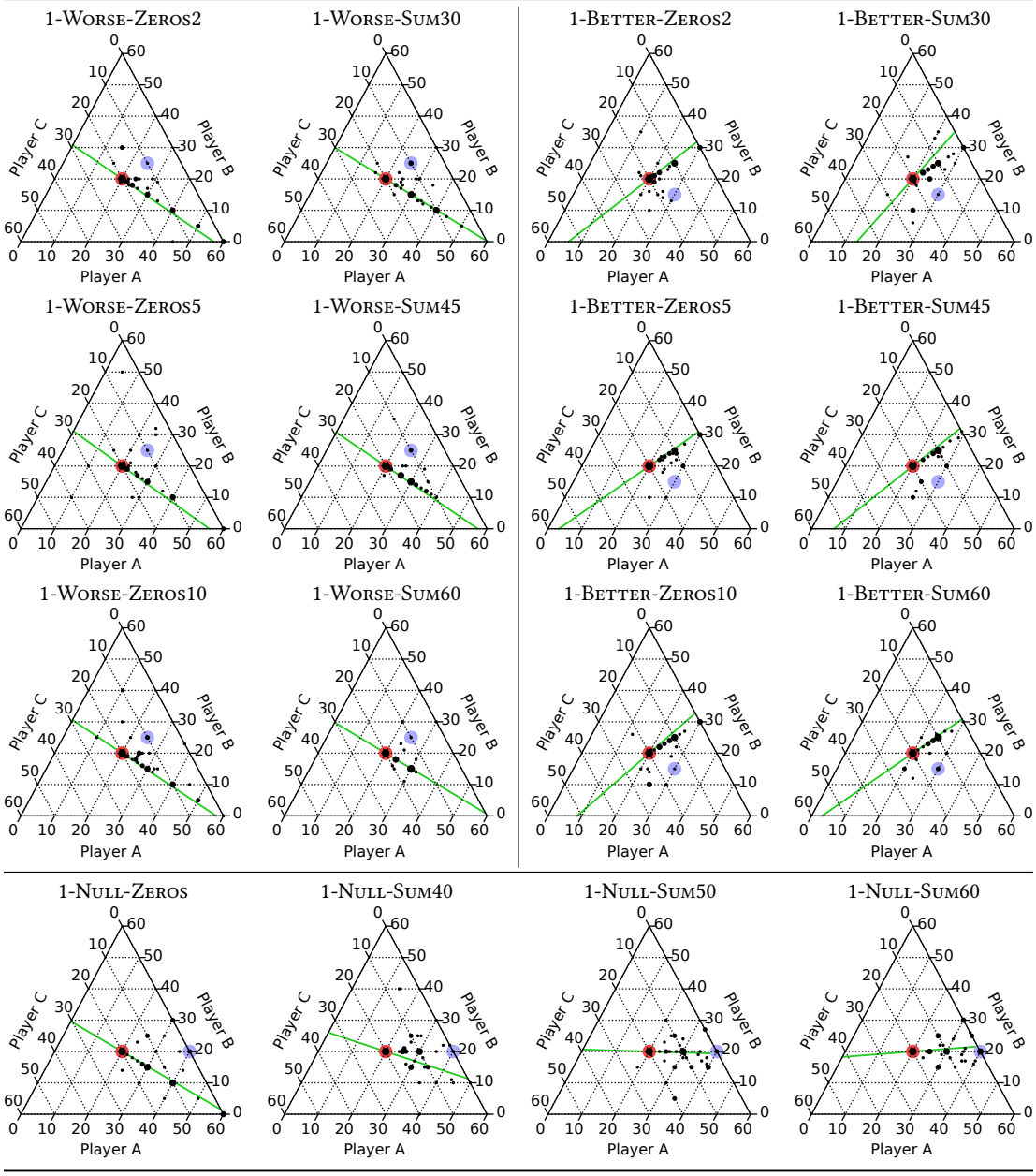

Figure 3: The rewards that participants submitted for each game in experiment 2. In each plot, $ED(f)$ is circled in dark red, and $Sh(f)$ is circled in light blue. Green lines indicate the direction of the main PCA component.

In these games, we found that 455/525 (86.7%) of the reward divisions obeyed symmetry, with 58/75 (77.3%) participants selecting no more than one division violating symmetry. We formalized this check by using paired Wilcoxon signed-rank tests to test whether the two symmetric players received different rewards. We found no significant differences between these rewards in any of these six games (all $p > 0.1$). Together, these results provide strong evidence that the symmetry axiom is consistent with our data.

Null Players: In all four of the 1-Null games in Experiment 2, player C is a null player. It is clear from Figure 3 that most participants gave a positive reward to player C, breaking the null player axiom. To quantify this behaviour, Figure 5 shows the cumulative distribution of the rewards that people assigned to player C in these games. This plot shows that few participants satisfied the null player axiom, with only 11% of their reward divisions giving a reward of 0 to the null player. Instead, many of the participants

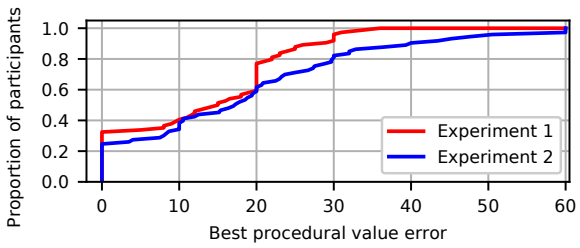

Figure 4: Empirical CDF of the errors made by the best-fitting procedural values.

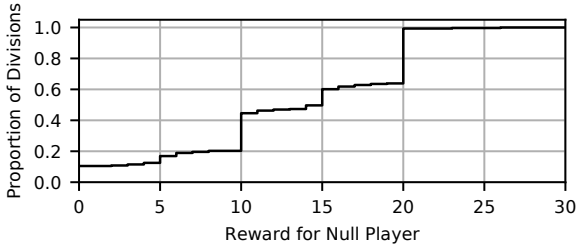

Figure 5: Empirical CDF of the rewards that were assigned to null players in Experiment 2.

tended to assign small, round rewards to the null player, with rewards of 5, 10, and 15 being most common, and a large jump at 20 due to the large number of equal divisions. While participants tend to recognize that null players contribute little to the group, they rarely go so far as to assign no reward to these null players.

Additivity: Several games in Experiment 2 are closely related. For instance, the 1-Worse-Sum30 and 1-Worse-Sum45 games only differ by the game

$$f(C) = \begin{cases} 5, & |C| = 1, \\ 0, & |C| \neq 1. \end{cases}$$

It is difficult to argue that any value other than $[0, 0, 0]$ is reasonable for this f: $f(N) = 0$, and all three players are symmetric. Thus, to satisfy additivity, participants must select the same rewards for all three of the 1-Worse-SumX games and for all three of the 1-Better-SumX games. We found that this was unlikely: only 258/444 (58.1%) of these comparisons succeeded, with only 26/74 (35.1%) participants violating the axiom at most once. Further, most of the successes (182/444 rewards and 21/74 participants) were explained by participants that always submitted equal divisions.

We made these comparisons rigorous using 6 within-subjects Friedman tests.[1] We found that the rewards for players A and C varied significantly in the 1-Worse games (both $p < 0.01$). We also found marginally significant results in the 1-Better games, with inconsistent rewards assigned to player A ($p = 0.08$) and player C ($p = 0.07$). These results imply that our data violates additivity.

[1]For instance, one test checked whether participants assigned the same rewards to player A in the 1-Worse-Sum30, 1-Worse-Sum45, and 1-Worse-Sum60 games.

5.3 Alternative Axioms for Reward Divisions

Knowing that people's reward divisions do not obey additivity leaves us with an enormous space of values: after removing additivity, any efficient reward division is acceptable in games with no symmetric or null players. Searching through this space for values that are a close match to people's rewards is a hopeless task.

We argue that the approach of testing alternative axioms against experimental datasets is a powerful methodological tool that can help structure this search. These alternative axioms are commonly proposed for the sake of characterizing existing solution concepts. For example, Young [40] famously showed that Shapley's null player and additivity axioms can be replaced with a *strong monotonicity* property. We propose using these axioms for a different purpose: rather than recovering existing values, we intend to use them to find additional structure on the people's reward divisions.

We illustrate this idea with the class of monotonicity axioms. While people must not obey strong monotonicity – if they did, they could not have violated the null player and additivity axioms – we consider two weaker alteratives. First, *local monotonicity* [6] puts an ordering on the players: if player i never makes a smaller marginal contribution to any coalition than player j, then i must not receive a smaller reward. Second, *coalitional monotonicity* [40] requires that, if a single coalition increases in value, its members's rewards cannot be decreased. Do people's rewards satisfy these two weaker axioms?

Local monotonicity: Every game tested in our experiments supports the hypotheses of local monotonicity in some capacity. For example, for a reward division r to be locally monotonic for any the 1-Worse games in Experiment 1, it must have $r_A = r_B \geq r_C$; in Experiment 2, it must have $r_A \geq r_C$ and $r_B \geq r_C$. We tested each participant's reward divisions to check if they met these conditions. In Experiment 1, we found that 734/825 (89.0%) reward divisions satisfied this property, with 59/74 participants (76.0%) submitting no more than one division violating local monotonicity. Experiment 2 was similar, with 1203/1258 (95.6%) reward divisions and 59/74 (79.6%) participants. These results indicate that local monotonicity is a reasonable description of people's rewards.

Coalitional monotonicity: The conditions of coalitional monotonicity hold for two pairs of games in Experiment 1. Between 1-Better-Solo and 1-Worse-Solo, only the value of player B's solo coalition is increased; 73/75 (97.3%) of participants gave player B a weakly higher reward in the latter. Then, 1-Worse-Pair and 1-Better-Pair only differ in the value of the coalition $\{A, C\}$, and 58/75 (77.3%) participants gave weakly higher rewards to both of these players. These results appear promising, but more work is required to make this evaluation rigorous.

While these two examples illustrate a first step toward testing alternative axioms, there are others that our data cannot validate. For some of these axioms, we would only need to adjust the coalitions' values. *Aggregate monotonicity* [40] and *weak monotonicity* [37] involve comparisons between games with different grand coalition values; our games all used a grand coalition value of 60. Others require more significant changes to the games. For example, there are a number of *consistency* axioms [37] that relate reward divisions between games with different numbers of players.

We note that the choice of interface is important when testing these axioms. Displaying the games in a tabular format worked well for relatively small and simple games, but might become unwieldy with more than 3 players. To test axioms in a wider variety of games, it might be necessary to design a new interface, taking ideas from MC-net [17] or coalitional skill game [2] representations of cooperative games. Studying how these alternative displays affect people's reward divisions is an interesting direction for future work.

5.4 A Data-Driven Axiomatic Approach

While the focus of our experiments was on cooperative games, the same high-level methods can be applied more broadly to fair algorithms in other domains. One exemplary problem is fair division, which also has a rich body of axiomatic work. For these problems, our methods can help to close the gap between axiomatic solutions and people's opinions. This process requires two key ingredients: a source of *data* and the *ability to test axioms*.

The first requirement is a source of data about people's opinions. One potential source for this data is through controlled experiments, as in our work. Controlling the scenarios can make it easier to design specific situations that allow axioms to be tested. Another is to consider in-the-wild applications of fair algorithms, giving scenarios that may be more representative of fair decisions in realistic situations. In either case, participants can reveal their opinions in several ways. As in our experiment, they might suggest fair outcomes [23]; they could also give fairness ratings by scoring individual outcomes [9, 11] or through pairwise comparisons [10, 26].

The second requirement is a method for testing whether people's opinions align with axioms' predictions. The simplest way to do this is to count how often people violate an axiom when its premises hold, as we did for each of our tests. However, a more thorough method is to quantify how drastically each axiom has been violated. We used this approach for the null player axiom, checking how much reward each participant gave to the null players. In general, it may be valuable to make this type of test more rigorous. For cooperative games, Aguiar et al.'s Shapley error decomposition [1] quantifies how drastically a reward division breaks each of the four axioms. Analogous tools could apply to other domains.

It might appear unsatisfying that this approach will not lead to a unique fair solution, but this limitation is necessary: our data showed that people have a range of opinions on fair outcomes, so we cannot hope to predict a unique output. This variety of opinions is consistent with Lee et al.'s findings in fair division [23, 24]. Furthermore, placing structure on the set of fair outcomes is sometimes more important than choosing one decision from this set. In "algorithm-in-the-loop" decisions, this final choice is left in human hands [12]. Small groups can also use these ranges as a starting point for discussions [23]. Other computational applications do not require an exact value, either: Freedman et al. [10] used fairness perceptions only to break ties between equally feasible outcomes.

Finally, Procaccia argues that a crucial strength of the axiomatic method is its ability to produce natural-language explanations of outcomes [34]. This property is important, as explaining standards and outcomes is one of the core requirements for designing a procedurally fair algorithm [24]. Identifying alternative axioms that match the data can help to provide these explanations for both human and algorithmic choices. In particular, axioms that are punctual [3] – putting conditions on a single outcome, rather than on the relationship between several outcomes – can often be translated into these simple explanations. Thus, validating axioms with data unites two notions of fairness: it lets us create algorithms that are aligned with stakeholders' notions of distributive fairness without losing the ability to explain the principles behind these outcomes.

6 VALIDITY AND LIMITATIONS

Before we conclude, we address the validity of our dataset. Our participants were workers on Mechanical Turk, who often focus on completing their tasks quickly. It is difficult for us to measure workers' comprehension or effort beyond our simple tutorial and filters. Despite these potential issues, we still see significant value in our results: our data shows clear trends, indicating the consistency of these workers across many games. Thus, while crowdsourcing these experiments might explain the high rate of equal divisions in many games, we believe that our data successfully captures human heuristics for these games. Future work could study how people's perceptions of fair reward divisions differ between populations.

We also note that it is difficult to say how our results depend on the framing of the game. In our experiments, we gave a story of three people playing a video game online. One participant explicitly mentioned this story in a post-survey, stating that it is most common for parties to split their loot evenly in these games, regardless of the members' contributions. It would be interesting to study how this behaviour would change if the video game setting was replaced with a merger negotiation between several companies, or if participants divided losses or costs instead of rewards. Reframing the games in this way might induce more calculated, rational outcomes, capturing fairness in a different setting.

7 CONCLUSION

In this paper, we studied how humans divide rewards as impartial decision makers in cooperative games. Our results showed that the values of the single-player coalitions in these games, which have typically been fixed at zero in previous work, play an important role in people's reward division decisions. In many situations, the values of these single-player coalitions appear to take precedence over the two-player coalitions. We used our data to show that people respect the symmetry axiom, but not the null player or additivity axioms that are used to characterize the Shapley value.

In light of these results, we argue that our general methodology of testing axioms against experimental data is a promising research direction for designing practical, fair algorithms. Even when these axioms are not used to derive a unique outcome, they help to place structure on the set of fair outcomes. In some settings, this structure is all that is needed to support a human decision. In others, it serves as a first step toward combining data-driven methods with the benefits of the axiomatic approach in designing fair algorithms.

ACKNOWLEDGMENTS

Thanks to Edith Law for funding the experiments and to the anonymous reviewers for their helpful comments. We acknowledge the support of the Ontario Graduate Scholarship and the NSERC CGS-D scholarship.

REFERENCES

[1] Victor H. Aguiar, Roland Pongou, and Jean-Baptiste Tondji. 2018. A non-parametric approach to testing the axioms of the Shapley value with limited data. *Games and Economic Behavior* 111 (Sep. 2018), 41–63. DOI: http://dx.doi.org/10.1016/J.GEB.2018.06.003

[2] Yoram Bachrach and Jeffrey Rosenschein. 2008. Coalitional skill games. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*. 1023–1030.

[3] Sylvain Béal, Eric Rémila, and Philippe Solal. 2015. A decomposition of the space of TU-games using addition and transfer invariance. *Discrete Applied Mathematics* 184 (Mar. 2015), 1–13. DOI: http://dx.doi.org/10.1016/J.DAM.2014.12.019

[4] Gary E. Bolton, Kalyan Chatterjee, and Kathleen L. McGinn. 2003. How communication links influence coalition bargaining: A laboratory investigation. *Management Science* 49, 5 (May 2003), 583–598. DOI: http://dx.doi.org/10.1287/mnsc.49.5.583.15148

[5] Colin Camerer. 2003. *Behavioural game theory: Experiments in strategic interaction*. Princeton Unversity Press.

[6] André Casajus and Frank Huettner. 2013. Null players, solidarity, and the egalitarian Shapley values. *Journal of Mathematical Economics* 49, 1 (jan 2013), 58–61. DOI: http://dx.doi.org/10.1016/J.JMATECO.2012.09.008

[7] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. 2011. Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 6 (2011), 1–168.

[8] Geoffroy De Clippel and Kareen Rozen. 2013. Fairness through the lens of cooperative game theory: An experimental approach. (2013). http://cowles.econ.yale.edu/

[9] Greg d'Eon, Joslin Goh, Kate Larson, and Edith Law. 2019. Paying Crowd Workers for Collaborative Work. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 125 (Nov. 2019), 24 pages. DOI: http://dx.doi.org/10.1145/3359227

[10] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P Dickerson, and Vincent Conitzer. 2018. Adapting a kidney exchange algorithm to align with human values. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[11] Ya'akov (Kobi) Gal, Moshe Mash, Ariel D. Procaccia, and Yair Zick. 2016. Which is the fairest (rent division) of them all?. In *Proceedings of the 2016 ACM Conference on Economics and Computation - EC '16*. ACM Press, New York, New York, USA, 67–84. DOI: http://dx.doi.org/10.1145/2940716.2940724

[12] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (Nov. 2019), 24 pages. DOI: http://dx.doi.org/10.1145/3359152

[13] Yoram Halevy. 2007. Ellsberg Revisited: An Experimental Study. *Econometrica* 75, 2 (2007), 503–536. DOI: http://dx.doi.org/10.1111/j.1468-0262.2006.00755.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2006.00755.x

[14] Jason S Hartford, James R Wright, and Kevin Leyton-Brown. 2016. Deep Learning for Predicting Human Strategic Behavior. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 2424–2432. http://papers.nips.cc/paper/6509-deep-learning-for-predicting-human-strategic-behavior.pdf

[15] Dorothea K. Herreiner and Clemens Puppe. 2010. Inequality aversion and efficiency with ordinal and cardinal social preferences—An experimental study. *Journal of Economic Behavior & Organization* 76, 2 (2010), 238 – 253. DOI: http://dx.doi.org/https://doi.org/10.1016/j.jebo.2010.06.002

[16] Dorothea K. Herreiner and Clemens D Puppe. 2009. Envy freeness in experimental fair division problems. *Theory and decision* 67, 1 (2009), 65–100.

[17] Samuel Ieong and Yoav Shoham. 2005. Marginal contribution nets: A compact representation scheme for coalitional games. In *Proceedings of the 6th ACM Conference on Electronic Commerce*. ACM, 193–202.

[18] Reinoud Joosten. 1996. *Dynamics, equilibria, and values*. Ph.D. Dissertation. Maastricht University.

[19] James P. Kahan and Amnon Rapoport. 1977. When you don't need to join: The effects of guaranteed payoffs on bargaining in three-person cooperative games. *Theory and Decision* 8, 2 (1977), 97–127.

[20] James P. Kahan and Amnon Rapoport. 1984. *Theories of coalition formation*. Psychology Press.

[21] Anson Kahng, Min Kyung Lee, Ritesh Noothigattu, Ariel Procaccia, and Christos-Alexandros Psomas. 2019. Statistical Foundations of Virtual Democracy. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 3173–3182. http://proceedings.mlr.press/v97/kahng19a.html

[22] Gerhard K. Kalisch, John W. Milnor, John F. Nash, and Evan D. Nering. 1954. Some experimental n-person games. In *Decision Processes*, Robert M. Thrall, Clyde H. Coombs, and Robert L. Davis (Eds.). 301–327.

[23] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1035–1048. DOI: http://dx.doi.org/10.1145/2998181.2998230

[24] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 182 (Nov. 2019), 26 pages. DOI: http://dx.doi.org/10.1145/3359284

[25] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article Article 182 (Nov. 2019), 26 pages. DOI: http://dx.doi.org/10.1145/3359284

[26] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 181 (Nov. 2019), 35 pages. DOI: http://dx.doi.org/10.1145/3359283

[27] Joshua Lewis, Margareta Ackerman, and Virginia de Sa. 2012. Human cluster evaluation and formal quality measures: A comparative study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34.

[28] Marcin Malawski. 2013. Procedural values for cooperative games. *International Journal of Game Theory* 42, 1 (Feb. 2013), 305–324. DOI: http://dx.doi.org/10.1007/s00182-012-0361-7

[29] Michael Maschler. 1992. The bargaining set, kernel, and nucleolus. In *Handbook of game theory with economic applications*, Robert J. Aumann and Sergiu Hart (Eds.). Vol. 1. Elsevier, 591–667.

[30] John F. Nash, Rosemarie Nagel, Axel Ockenfels, and Reinhard Selten. 2012. The agencies method for coalition formation in experimental games. *Proceedings of the National Academy of Sciences* 109, 50 (2012), 20358–20363. DOI: http://dx.doi.org/10.1073/pnas.1216361109

[31] Ritesh Noothigattu, Snehalkumar S Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D Procaccia. 2018. A voting-based system for ethical decision making. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[32] Andrzej S. Nowak and Tadeusz Radzik. 1994. A solidarity value for n-person transferable utility games. *International Journal of Game Theory* 23, 1 (Mar 1994), 43–48. DOI: http://dx.doi.org/10.1007/BF01242845

[33] Andrzej S Nowak and Tadeusz Radzik. 1996. On convex combinations of two values. *Applicationes Mathematicae* 24, 1 (1996), 47–56.

[34] Ariel Procaccia. 2020. Axioms Should Explain Solutions. In *Future of Economic Design*, Jean-François Laslier, Hervé Moulin, M. Remzi Sanver, and William S. Zwicker (Eds.). Springer, Forthcoming.

[35] Tadeusz Radzik and Theo Driessen. 2013. On a family of values for TU-games generalizing the Shapley value. *Mathematical Social Sciences* 65, 2 (mar 2013), 105–111. DOI: http://dx.doi.org/10.1016/J.MATHSOCSCI.2012.10.002

[36] Lloyd S. Shapley. 1953. A value for n-person games. In *Contributions to the Theory of Games* (2nd ed.), H. Kuhn and A.W. Tucker (Eds.). Princeton Unversity Press, Princeton, 307–317. https://pdfs.semanticscholar.org/ba28/dafc77ec548d1f7d7af09c9470ae7b6752dd.pdf

[37] René van den Brink, Yukihiko Funaki, and Yuan Ju. 2013. Reconciling marginalism with egalitarianism: consistency, monotonicity, and implementation of egalitarian Shapley values. *Social Choice and Welfare* 40, 3 (2013), 693–714.

[38] Michael A. Williams. 1988. An Empirical Test of Cooperative Game Solution Concepts. *Behavioral science* 33, 3 (Jul 01 1988), 224. http://ezproxy.library.ubc.ca/login?url=https://search.proquest.com/docview/1301274414?accountid=14656 Last updated - 2013-02-24.

[39] James R. Wright and Kevin Leyton-Brown. 2010. Beyond equilibrium: Predicting human behavior in normal-form games. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)* (2010), 901–907. DOI: http://dx.doi.org/978-1-57735-463-5

[40] H. P. Young. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory* 14, 2 (Jun 1985), 65–72. DOI: http://dx.doi.org/10.1007/BF01769885

[41] Yair Zick, Kobi Gal, Yoram Bachrach, and Moshe Mash. 2017. How to form winning coalitions in mixed human-computer settings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 465–471. DOI: http://dx.doi.org/10.24963/ijcai.2017/66