Validating Generic Metrics of Fairness in Game-based Resource Allocation Scenarios with Crowdsourced Annotations

Corrado Grappiolo[1], Héctor P. Martínez[2], Georgios N. Yannakakis[2]

[1] Center for Computer Games Research, IT University of Copenhagen, Denmark
cogr@itu.dk
[2] Institute of Digital Games, University of Malta, Malta
{hector.p.martinez, georgios.yannakakis}@um.edu.mt

Abstract. Being able to effectively measure the notion of *fairness* is of vital importance as it can provide insight into the formation and evolution of complex patterns and phenomena, such as social preferences, collaboration, group structures and social conflicts. This paper presents a comparative study for quantitatively modelling the notion of *fairness* in *one-to-many* resource allocation scenarios — i.e. one *provider* agent has to allocate resources to multiple *receiver* agents. For this purpose, we investigate the efficacy of six metrics and cross-validate them on crowdsourced human ranks of fairness annotated through a computer game implementation of the one-to-many resource allocation scenario. Four of the fairness metrics examined are well-established metrics of data dispersion, namely standard deviation, normalised entropy, the Gini coefficient and the fairness index. The fifth metric, proposed by the authors, is an ad-hoc context-based measure which is based on key aspects of distribution strategies. The sixth metric, finally, is machine learned via ranking support vector machines (SVMs) on the crowdsourced human perceptions of fairness. Results suggest that all ad-hoc designed metrics correlate well with the human notion of fairness, and the context-based metrics we propose appear to have a predictability advantage over the other ad-hoc metrics. On the other hand, the normalised entropy and fairness index metrics appear to be the most expressive and generic for measuring fairness for the scenario adopted in this study and beyond. The SVM model can automatically model fairness more accurately than any ad-hoc metric examined (with an accuracy of 81.86%) but it is limited by its expressivity and generalisability.

Keywords: Fairness, Social Preference, Resource Allocation, Dictator Game, Crowdsourcing, Preference Learning, Support Vector Machines, Feature Selection, Genetic Algorithms.

1 Introduction

The control and influence of virtual or artificial societies is a highly complex task, in part, due to the difficulty of predicting the reaction and evolution of the

population to dynamic elements or changes. To monitor and predict evolution of a society such as a complex adaptive system [37] one needs to monitor the behaviour of single individual agents within the society. The behaviour of the individuals collectively generates complex (and emergent) global dynamics and phenomena (e.g. friendship networks); these in turn affect the individuals, who will adapt their behaviour autonomously creating a loop that regulates how the society evolves. For example, the level of collaboration among the individuals in a local community can be utilised to develop a plan that allows for the integration of different ethnic group identities [53]. Alternatively, in a virtual society (e.g. a serious multiplayer game) collaboration could be monitored to effectively teach soft social skills [61].

In this paper we investigate a number of metrics to measure the *fairness* that one individual agent manifests depending on its interactions with other agents. We claim that fairness is a key feature characterising the interactions, which could bring further insight into the ongoing complex dynamics. The computational models of fairness derived can assist in the inference of social preferences, collaboration, group structures and consequently social conflicts within artificial societies and complex networks. For instance, an individual who treats individuals differently might suggest the existence of preference; this would also have implications for the reciprocity of such treatment, and possibly for altruism, collaboration, and group identities [9, 15].

We restrict our investigation to a virtual environment featuring a resource allocation scenario, which can allow us to simulate well and isolate scenarios which can be encountered in real-life situations. In the *one-to-many* scenario examined, an agent has to collect and share several resources among a population of agents divided into two visible groups. Although the scenario under investigation has features which are common to well-studied social dilemmas, such as the *other-other* game [6, 35], the *dictator game* [3] and, partly, the *ultimatum game* [9, 15], the perspective we take and the interpretation we give to the term "fairness" differs from those adopted in game theory studies. In particular, the metrics we investigate do not aim to model the *utility* of the actions performed (inequity aversion) [4, 16, 17, 48], as we rather focus on identifying measures of equality of treatment towards multiple individuals independently of the provider's payoff. Furthermore, we are not making the canonical game theoretical assumption that the individuals considered in our scenarios are merely greedy [15], hence, we do not aim to investigate how the observed behaviours differ from some reference value (e.g. Nash equilibrium) [15]. In other words, we aim to investigate a property of the interactions which is complementary to the concepts of social preference's "altruism" and "reciprocity" [15]. Our concept of fairness assumes that individuals behave "well" (i.e. they are fair) within a subset of individuals of the population only: fairness aims to provide measures capable of identifying whether this inequality is occurring.

In this paper we test and compare four simple standard dispersion measures — namely *standard deviation*, *normalised entropy* [57], the *Gini coefficient* [41], and the *fairness index* [29] — that approximate fairness based on the levels

of satisfaction of resource acquisition manifested by the receiver individuals. We also propose a new, ad-hoc metric — hereafter called *temporal group-based* fairness — which is based on single interactions (resource allocations) and the ability of an individual agent to identify itself as part of a group [9]. We validate the five metrics against the reports of an online crowdsourcing survey in which human participants ranked the level of fairness in a wide variety of resource allocation scenarios. Additionally, we propose a sixth, data-driven modelling approach, in which genetic-feature selection combined with ranking Support Vector Machines (SVMs) [30] extract the most relevant context-based features of the scenario and infer a non-linear mapping between scenario attributes and the crowdsourced notions of fairness. Results obtained show that all ad-hoc metrics are highly consistent with the human notion of fairness as obtained from the crowdsourced data; the temporal group-based metric proposed, however, outperforms the other ad-hoc metrics in this scenario. Finally, the rank SVM, manages to produce a non-linear model that reaches 81.86% of accuracy in predicting fairness and proves to be the most consistent with the human notion of fairness.

This study is novel as, to the best of the authors' knowledge, there has not been any prior attempt to design, compare and cross-validate metrics of fairness against crowdsourced annotations of fairness. Moreover, this paper builds upon and extends the metric validation method proposed in [57] by introducing rank-based crowdsourcing as a tool for annotating complex social dynamics such as fairness. Clearly, our crowdsourced cross-validation study would strengthen earlier work, which used the metrics we considered [22–24, 41]; moreover, the proposed metrics open a plethora of new applications. For instance, they can be used as a mechanism for evaluating and promoting fairness in educational (serious) games [22] and, similarly, cooperation/altruism [7]; they could also be used as metrics to understand and evaluate collaborative games [12, 49, 54] — under the collective intelligence perspective — together with, or as an alternative to, the use of expert knowledge [50]. They could be used in fields not strictly linked to computer games, such as social network analysis [13, 45] or, even further, as tools aimed to understand the goodness of the partitioning of (social) networks into community structures [20, 36]. Moreover, the combination of crowdsourcing with game-based scenarios introduces a first insight for categorising discrepancies between experiments conducted on artificial societies [2, 14] and those on human participants [15, 25]. Additionally, the metrics could also be embedded within such artificial agents, e.g. together with inequity aversion [10, 56], in order to generate *believable, human-like* adaptive artificial societies which manifest both greedy and group-based behaviours [47], allowing thus to further investigate the notion of evolution of (group-based) collaboration [6, 25, 26, 44].

The remainder of the paper is organised as follows: Section 2 lists some of the most relevant work to this study. Section 3 describes, formally, the one-to-many resource allocation interaction scenario used in our research; it illustrates how the scenario is implemented in a game-based virtual environment, and finally it delineates the design of the crowdsourcing protocol. Section 4 and Section 5 describe, respectively, the application of the ad-hoc generic metrics of fairness

and the use of SVMs for the one-to-many scenario. Section 6 presents the key results of this study. Section 7 discusses the limitations and the generalisability of the key findings and Section 8 concludes the paper.

2 Related Work

This study shares aspects of two main research fields: evolutionary game theory and game-based modelling. The studies on the *ultimatum* and the *dictator* game — linked to the resource allocation game we investigate in this paper — are numerous (see [3, 16, 17] and [33] among others). The aforementioned studies hold a pure *game theoretical* perspective: they aim to understand why the experiments conducted on human participants deviate from the Nash equilibrium — which assumes the existence of solely greedy individuals — towards more altruistic behaviours. The focus of our research, instead, is to provide metrics capable of capturing whether the individuals of a population are manifesting different levels of altruism, depending on who they interact with, as a key aspect of complex dynamics, such as the existence of group structures and group identities [9].

Possibly, even closer to our research aim, there exist studies focused on understanding the influence that group identities have on the behaviours of the participants of game theory experiments (see [1, 5, 6] and [34] among others); however, our work differs from these in its investigation of non-payoff (utility) based measures. Similarly, Marzo et al. [42] investigated how the existence of friendship structures affects the levels of altruism of human participants in *Colored Trails* [18], a computer game with similar mechanics to the ultimatum game. Their work puts the emphasis on describing how social relationships influence the behaviour of the participants in their experiment, which is in line with our assumptions about fairness; however, our work goes a step further by assuming the existence of such behaviours within the population and aiming to provide metrics capable of identifying them.

Studies on social preferences have benefited from the relatively novel agent-based modelling approach for simulating complex dynamics in artificial societies [14, 26]. This approach has a close, if not exact, relationship with the research fields of evolutionary dynamics, collective behaviour, and simulation studies, which aim to better understand how collaboration emerges from the repeated interactions among the agents [2, 28, 44]. Although our global aim is the understanding of the evolution of collaboration and group structures, the study presented in this paper does not rely on simulations of artificial societies and on the search for behavioural rules embedded in the agents.

Relevant metrics to fairness, such as balance [38] and symmetry [41] — key concepts in computer games research and design — have been constructed as they intuitively have an impact on both the entertainment and learning goals of the game; however, we argue that the notion of fairness is both different and more complex than symmetry and balance as it factors in social, behavioural and relationship effects. The context-based metrics proposed in this paper follow both a top-down *expert-driven* and a bottom-up *data-driven* [60] approach for cap-

turing fairness inspired by methodologies for deriving accurate user-interaction metrics in games [57]. While our expert-driven metrics are directly designed on key aspect of the one-to-many scenario in use, the evaluation and construction of the data-driven metric is based on game-based crowdsourced annotations.

The use of crowdsourcing and the correlation analysis we performed has been used in game artificial intelligence research, though centred on affect modelling in single player game environments [51, 57, 59]. Our one-to-many interaction scenario, however, provides the basis for the investigation of multi-agent and multiplayer scenarios [21, 23, 24].

The Fairness Index (FI) we decided to consider in this study [29] belongs to the vast research field of fairness of resource allocation in information networks (consider [11, 32, 43, 46, 52] and [55] as a non-exhaustive list). Generally speaking, the standard problem regards the *fair* allocation of resources (e.g. CPU and memory) to a set of requesting units (e.g. jobs) in order to e.g. maximise the overall system's *utility*. Although our scenario and aim is more targeted towards social and economical contexts, the positive correlation of FI with the perceived notion of fairness gathered in our crowdsourcing survey suggests that FI is relevant for our purposes.

3 Fairness Scenario and Crowdsourcing

This section presents the one-to-many resource allocation scenario on which we base our fairness modelling studies and describes the game-based virtual environment designed for encapsulating the key characteristics of the scenario. The section ends with a description of the protocol followed for crowdsourcing annotations of fairness through the virtual environment.

3.1 One-to-many Resource Allocation Scenario

Formally, the one-to-many resource allocation scenario investigated in this paper can be described by the following tuple:

$$\text{scenario} = (P, R, A, G, g, v, s) \tag{1}$$

where:

- P is the population (society) of $n{+}1$ individuals (agents), $P = \{a_0, a_1, \ldots, a_n\}$.
- R is the set of m resources $R = \{r_0, r_1, \ldots, r_{m-1}\}$.
- $A \subseteq R$ is the temporally ordered sequence of resources allocated by the provider agent, $A = \{r_i^{t=1}, r_j^{t=2}, \ldots\}$.
- G is the set of group structures present in P; G constitutes a partition of P.
- $g : P \to G$ is the *group identity function* mapping each agent in P to *one* group structure in G. The group identity of the provider agent is unknown to the observer of the distribution strategy; for simplicity, we will refer to the single group identity $g(a_i)$ as g_i.

- $v : R \rightarrow (0, 1]$ is the *resource value function* which assigns a value to each resource in R. For simplicity we will refer to the resource value of resource r_j as v_j.
- $s : P \rightarrow [0, 1]$ is the *receiver satisfaction function* mapping each receiver agent in P to a satisfaction value between 0 and 1. For simplicity we will refer to the satisfaction value of a_i, as s_i.

All the agents a_i have their satisfaction values s_i decreasing over time by a constant value δ, $s_i \leftarrow max(s_i - \delta, 0)$. An agent a_i receiving a resource r_j will have its satisfaction value updated as follows: $s_i \leftarrow min(s_i + v_j, 1)$. Intuitively, s_i gives an indication of the amount of resources acquired by a_i.

One agent takes the role of the provider agent (we will refer to it as $a_p \in P$), and the remaining n individuals take the role of the receiver agents. The provider agent has the duty to fill up A by allocating $0 \leq |A| \leq m$ resources among the n receiver agents in $|A|$ steps (i.e. one resource at a time, the provider agent may not distribute any resources, a receiver agent may obtain more than one resource). Each receiver agent a_i has a goal which corresponds to the maximisation of its s_i. However, the receiver agents cannot acquire the resources by themselves; they will only acquire the resources distributed by the provider agent.

The features describing the scenario allow the provider agent to adopt many different strategies to define A, which would *suggest*, to an external observer, different levels of fairness towards the receiver agents and, possibly, also receiver groups.

3.2 The Resource Allocation Game

To obtain a test-bed to evaluate the ad-hoc fairness metrics, and machine learning alternative data-driven metric, the one-to-many resource allocation scenario described in Section 3.1 was implemented as a three-dimensional (3D), single player game-based virtual environment; a virtual camera follows the movement of the provider agent, creating a third-person perspective of the gameplay. The provider agent finds himself in the environment together with receiver agents and resources; his duty is to allocate the resources among the receiver agents within a time limit by carrying only one resource at a time.

The satisfaction values of each receiver agent, which decrease constantly over time, can be inferred by observing their yellow hats, which range in transparency levels: the more transparent, the lower the satisfaction. When the receiver agent acquires a resource, its satisfaction level is increased (i.e. its yellow hat gets more opaque) by the value of the resource obtained. The resources are represented as purple balls and their values are represented by their size: the smaller the ball, the smaller its value provided.

The receiver agents have a coloured body which represents the group they belong to (group identity). For the study presented in this paper, we examine the presence of only two group structures, namely the red and the blue group. A single distribution scenario (Equation 1) is interpreted, in our game-based

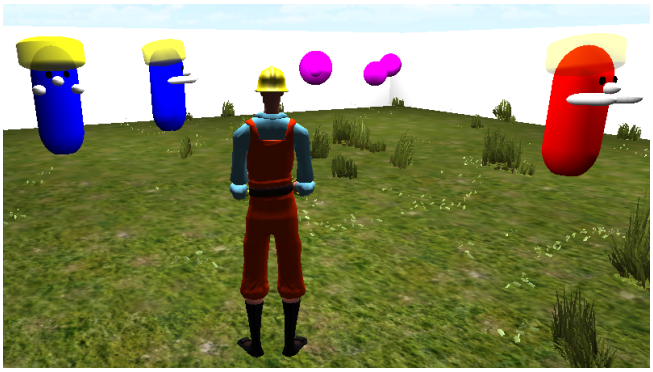

Fig. 1. A snapshot of the resource allocation game: 3 receiver agents (2 belonging to the blue group and 1 belonging to the red group) and their satisfaction levels (yellow hats), 4 resources of different sizes and the provider agent are depicted.

environment, as a game level. A snapshot of the resource allocation game can be observed in Figure 1.

The decision to consider only two groups in our study is motivated by several arguments, both theoretical and practical. First, the *red* and *blue* groups can be easily linked to the *in-group* vs. *out-group* dichotomy [6, 9]: the detection of unfair treatments — even across the whole population, i.e. independently on the agent colours/groups — could be augmented by further detection of the targeted subset of the population which received the most generous treatments, and hence the subsequent identification of the colour/group the provider agent belongs to [23, 24]. Second, the fact that some of the most relevant studies on group behaviour and inequity aversion in human-centred experiments [6, 34, 42] are based on two groups would allow for a possibly easier bridging between the two different interpretations given to fairness. Finally, we did not want to introduce any further complexity to the game-based scenario, which could bring difficulties with respect to the perception of group-based unfair treatments (see Subsection 4.5), given the novelty of the approach we decided to adopt.

3.3 Crowdsourcing Experimental Protocol

The game-based virtual environment was then used in a crowdsourcing survey. In the survey, participants watched and annotated a sequence of *strategy videos*, which reproduce resource distribution tasks (strategies) of our game. We have relied on the ability of participants to compare and rank gameplay strategies given in pairs (pairwise preference scheme) rather than having participants rating the strategies, for a number of advantages including the elimination of the subjective notion of scaling and effects related to reporting order [58].

Table 1. Initial conditions of the four different scenarios of the survey. The s_i values are partitioned according to the two *red* and *blue* group identities.

Scenario	n_{red}	n_{blue}	m	s_{red}	s_{blue}	v_{R}
2 receivers	1	1	2	0.1	0.5	0.5, 0.9
3 receivers	2	1	4	0.1, 0.5	0.5	0.1, 0.3, 0.5, 0.9
4 receivers	2	2	3	1, 0.8	0.2, 0.1	0.8, 0.4, 0.2
5 receivers	3	2	4	0.1, 0.3, 0.5	0.7, 0.9	0.3, 0.5, 0.7, 0.9

Six different strategies were recorded for four different scenarios — consisting of 2, 3, 4 and 5 receiver agents — resulting in 24 strategy videos in total[3]. The set of game-playing strategy videos used in the survey was defined in order to show resource distribution strategies as different from each other as possible. With respect to the game features we have described in Section 3.2, each game-playing strategy is limited to 30 seconds. Table 1 illustrates the key features of the four different scenarios designed: number of red receiver agents ("n_{red}" column), number of blue receiver agents ("n_{blue}" column), number of resources ("m" column), initial satisfaction values of the receiver agents ("s_{red}" and "s_{blue}" columns) and value of the resources ("v_{R}" column). As can be noted, the four scenarios present unbalanced initial conditions, both with respect to the single agents and to the average group satisfaction values.

The key protocol steps of the survey are as follows:

1. The participant fills in a demographic questionnaire (age, gender, experience in playing games, whether the participant has already taken part in the survey) and reads the instructions of the survey, which describe the resource allocation scenario and the ways to provide feedback about the strategies.
2. A pair of strategy videos, A and B, is selected and presented on the same screen (see Figure 2) — there is a 50% probability that the order of appearance (left or right) of two strategies is inverted. The pair of strategy videos presented to the participants belong to the same scenario (i.e. same number of receivers).
3. The participant watches the videos and provides feedback through a 4-alternative forced choice questionnaire (4-AFC), which asks the participant if the behaviour of the provider was: (a) more fair in video A; (b) more fair in video B; (c) equally fair in both video A and video B; (d) unfair in both video A and video B. In order to allow the participant to fully grasp the distribution dynamics, each video can be paused at any time and replayed unlimitedly.
4. The participant can write additional comments via a free-response text box and is free to take part in another session of the survey, ensuring that he does not watch the same pair of strategy videos; otherwise the survey ends.

[3] The videos of the 24 Scenarios can be found on the following link: `http://itu.dk/people/cogr/FairnessExperiment/TCCI_index.php`

Fig. 2. A snapshot of the online survey

Considering that we have designed six strategies per scenario, and that we show pairs of different strategies of the same scenario, we get $C_{6,2} = 15$ possible combinations of video pairs by excluding repetitions and order of appearance on screen (left or right). Any possible order effects are minimised via the randomisation of the video ordering previously described.

For each of the four scenarios and the 15 possible combinations, we collected four pairwise preferences. Thus, in total, our analysis is based on $4 \cdot 15 \cdot 4 = 240$ strategy pairs that are labelled by our participants. The participants were gathered through advertisement of the experiment on the well-know Facebook[4] social network online platform, in order to achieve the most diverse audience. The number of unique participants out of the 240 reports are 141, 74% of which are male and 50.3% consider themselves gamers. The average age is 30.24 years and its standard deviation is 6.84. Furthermore, 77% of the participants declared themselves as being either not a gamer or occasional gamers (i.e. playing up to 2 hours per week), 11% declared playing games from 2 to 5 hours per week, 8% play from 5 to 10 hours per week, and 4% play for more than 10 hours per week.

4 Ad-Hoc Metrics of Fairness

This section first introduces the four generic metrics of dispersion of the data used in our study, given the particular scenario described by the tuple defined in Equation 1. The section ends with the definition of a new, ad-hoc designed fairness metric proposed by the authors, which is tightened to the scenario investigated.

[4] `http://www.facebook.com`

4.1 Standard Deviation

At the end of the distribution task, we calculate the average satisfaction of all receiver agents as follows:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} s_i \tag{2}$$

The standard deviation is therefore defined as follows:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (s_i - \mu)^2} \tag{3}$$

Since the upper bound of σ is 1 — both μ and s_i values range within the $[0, 1]$ interval — we can subtract σ from its maximum value in order to give it a positive connotation of fairness:

$$SD = 1 - \sigma \tag{4}$$

In presence of fairness, the s_i values are similar to each other, making σ close to zero and SD high. In contrast, in presence of unfair treatments, the s_i values are very different from each other, making σ high and SD low.

4.2 Normalised Entropy

The Normalised Entropy [57] is calculated at the end of each distribution task as follows:

$$NH = -\frac{1}{log(n)} \sum_{i=1}^{n} \hat{s}_i \log \hat{s}_i \tag{5}$$

where $\hat{s}_i$ represents the normalised satisfaction value of each receiver agents:

$$\hat{s}_i = \frac{s_i}{\sum_{j=1}^{n} s_j} \tag{6}$$

In presence of fairness, the $\hat{s}_i$ values are similar to each other, thus yielding a high normalised entropy. In contrast, in presence of unfair treatments, the $\hat{s}_i$ values are very different from each other, yielding low normalised entropy values.

4.3 Gini Coefficient

The Gini Coefficient [19] is a well-known measure of equality used in economic studies. Providing that, at the end of each level, the s_i values of the receiver agents are ordered by increasing values, that is:

$$s_i^k \leq s_j^{k+1}, \ i, j \in [1, n] \tag{7}$$

The calculation of the Gini Coefficient in our game-based scenario is done as follows:

$$GI = \frac{1}{n-1}\left(n+1-2\left(\frac{\sum_{k=1}^{n}(n+1-k)\,s^k}{\sum_{k=1}^{n}s^k}\right)\right) \tag{8}$$

Note that we have omitted the indices i in the representation for s^k for simplicity. GI lies within the $[0,1]$ interval: the lower the coefficient, the higher the fairness. Hence, similarly to the SD implementation, we will take into consideration the inverse of the Gini Coefficient:

$$GC = 1 - GI \tag{9}$$

4.4 Fairness Index

The Fairness Index [29] is a well-known measure of equality of resource allocation vastly used in information networks. The metric is calculated at the end of each distribution task as follows:

$$FI = \frac{\left(\sum_{i=1}^{n} s_i\right)^2}{n\sum_{i=1}^{n} s_i^2} \tag{10}$$

FI has value one in presence of completely fair treatments among the whole population, and it decreases as the disparity increases towards a subset of few individuals.

4.5 Temporal Group-based Metrics of Fairness

We here propose a new metric of fairness — which is not as generic as the four aforementioned metrics previously introduced — whose design is based on (and bounded by) the following three constraints (or criteria):

(C1) The group identities of the receiver agents are not hidden from the provider agent. The provider agent can be influenced by such information and therefore aim to deliver the resources based on the existing group structures, rather than just focusing on the receiver agents' satisfaction values.

(C2) The complex (and dynamic) structure of the scenario might have an impact on the willingness of the provider agent to be either fair or unfair. The metrics calculation should take into account the intermediate steps of the distribution task, rather than solely focusing on the calculation at the end of the scenario.

(C3) The initial configuration of each scenario might already depict an unfair distribution of s_i values. A provider agent, who is willing to maximise the satisfaction values of the whole population, should not avoid the delivery of resources.

In our game, which considers two group structures — the *red* and *blue* agents, see Section 3.2 — constraint (C1) is satisfied by the average satisfaction of the red and the blue groups:

$$\mu_{\text{red}} = \frac{1}{n_{\text{red}}} \sum_{i=0}^{n_{\text{red}}} s_i, \;\; \mu_{\text{blue}} = \frac{1}{n_{\text{blue}}} \sum_{i=0}^{n_{\text{blue}}} s_i \tag{11}$$

where, $n_{\text{red}} + n_{\text{blue}} = n$ is the number of receiver agents belonging to the red and the blue group, respectively. The between-group fairness is defined as the absolute value:

$$|\mu_{\text{red}} - \mu_{\text{blue}}| \tag{12}$$

Constraint (C2) is satisfied by averaging the between-group difference across the $0 < |A| \leq m$ distributions:

$$\frac{1}{|A|} \sum_{t=1}^{|A|} \left|\mu_{\text{red}}^t - \mu_{\text{blue}}^t\right| \tag{13}$$

Where μ_{red}^t and μ_{blue}^t represent, respectively, the between-group satisfaction values of the red and the blue groups after the t-th resource allocation. Finally, we consider the case $|A| = 0$ in order to satisfy (C3):

$$TGB = \begin{cases} 0 & \text{if } |A| = 0 \\ 1 - \frac{1}{|A|} \sum_{t=1}^{|A|} |\mu_{\text{red}}^t - \mu_{\text{blue}}^t| & \text{if } 0 < |A| \leq m \end{cases} \tag{14}$$

In presence of group-based fairness throughout the whole distribution task, the μ_{red} and μ_{blue} values are, on average, similar to each other, thus yielding high TGB values. In contrast, in presence of group-based unfair treatments throughout the whole distribution task, the μ_{red} and μ_{blue} are, on average, different from each other, generating low TGB values.

5 Data-Driven Approach: Preference Learning

We propose a data-driven method to model fairness as an alternative to the hand-crafted metrics defined in Section 4. This approach imposes minimal expert knowledge on the metrics which are instead shaped according to a specific dataset containing strategies with different levels of fairness compared by human judges. A data-driven technique, as opposed to hand-crafted metrics, can handle large amounts of user data and learn patterns that human experts may have overseen.

We use ranking Support Vector Machines (SVMs) [27, 31], a well-known method for learning non-linear estimators of user pairwise preferences such as those present in the experimental data used in this paper. This method requires that each input sample (i.e. strategy) is coded as a vector of numeric features given by the designer. To reduce the human bias that could be introduced in

Table 2. Full feature list extracted for each allocation video across the four scenarios. Bold feature numbers indicate the best extracted features by SVM's the genetic feature selection phase.

Feature Name	Description
n_{red}/n, n_{blue}/n	number of red/blue agents
$\lvert A\rvert/m$	number of deliveries
$n_{\text{red}}/3$, $n_{\text{blue}}/2$	number of red/blue agents divided by 3/2, i.e. their maximum number across all scenarios
$n_{\text{red}}/5$, $n_{\text{blue}}/5$	number of red/blue agents divided by 5, i.e. the maximum number of agents across all scenarios
$m/4$	number of resources divided by 4, i.e. their maximum number across all scenarios
$\lvert A\rvert/4$	number of deliveries divided by 4, i.e. the maximum number of possible deliveries across all scenarios
μ^0_{red}, μ^0_{blue}	initial average happiness of the red/blue agents
$\sum_{j=1}^{m} v_j$	average satisfaction value of the resources
v_j^t	for $t = 1, 2, 3, 4$, the value of the t-th resource delivered. If $\lvert A\rvert < t$ then the feature has value zero.
μ^t_{red}	for $t = 1, 2, 3, 4$, average happiness of the red agents after the t-th delivery. If the resource was delivered to the blue group, then $\mu^t_{\text{red}} = \mu^{t-1}_{\text{red}} - \delta$. If $\lvert A\rvert < t$ then the feature has value zero.
μ^t_{blue}	for $t = 1, 2, 3, 4$, average happiness of the blue agents after the t-th delivery. If the resource was delivered to the red group, then $\mu^t_{\text{blue}} = \mu^{t-1}_{\text{blue}} - \delta$. If $\lvert A\rvert < t$ then the feature has value zero.

this step, we use a large vector to describe each distribution strategy within a scenario and later reduce it by automatic feature selection. The remainder of this section describes the feature extraction process, the SVM, and the genetic-search feature selection algorithm used.

5.1 Fairness Strategy Feature Extraction

To capture most aspects of the scenario in a format suitable for data-driven modelling we extract a large number of features which could represent both its initial conditions and its intermediate ones. In total, we extracted 24 features, presented in Table 5.1.

5.2 Ranking Support Vector Machines

A Support Vector Machine is a binary classifier consisting of a linear combination of training vectors:

$$\mathbf{w} = \sum_{\mathbf{x_i} \in D} \alpha_i \phi(\mathbf{x_i}) \tag{15}$$

The hyper-plane defined by this combination separates the data samples $\mathbf{x_i}$ into two classes in projected space $\phi(X)$. In this paper we are not interested in a binary classifier but rather a function that maps data samples (strategies) into a real-valued variable (estimation of fairness); additionally, our data samples are associated with pairwise comparisons instead of discrete classes. Thus, we used a variant of SVMs known as ranking SVM [27, 31], which is tailored to problems with the same formulation. The model is still a linear combination of transformed training vectors which is trained by optimising the following problem:

$$\begin{aligned} \text{minimise:} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum \xi_i \\ \text{subject to:} \quad & \forall(\mathbf{x_P^i}, \mathbf{x_N^i}) \in D, \mathbf{w} \cdot (\phi(\mathbf{x_P^i}) - \phi(\mathbf{x_N^i})) \geq 1 - \xi_i \\ & \forall i \ \xi_i \geq 0 \end{aligned} \tag{16}$$

where D is the complete set of training samples, $(\mathbf{x_P^i}, \mathbf{x_N^i})$ are pairs of training samples such that the feature vector $\mathbf{x_P^i}$ was preferred (reported as *more fair* in this paper) over the feature vector $\mathbf{x_N^i}$, ξ_i are non-negative variables, C a weighting parameter, $\mathbf{w}$ the trained decision boundary and $||\mathbf{w}||$ its module.

Once $\mathbf{w}$ is trained, given a pair of samples $\{\mathbf{x_i}, \mathbf{x_j}\}$ the SVM predicts that $\mathbf{x_i}$ is preferred over $\mathbf{x_j}$ if $\mathbf{w}\cdot\phi(\mathbf{x_i}) > \mathbf{w}\cdot\phi(\mathbf{x_j})$; thus, $\mathbf{w}\cdot\phi(\mathbf{x})$ is a computational predictor of fairness. Although the SVM creates a linear separation, this is defined on the transformed space defined by ϕ, which yields more complex functions in the input space. As the transformation of the space can lead to an infinite number of dimensions [27], and also in order to reduce computational costs, the predictor is redefined in terms of the training examples and a kernel function:

$$\mathbf{w} \cdot \phi(\mathbf{x}) = \sum_{(\mathbf{x_P^i}, \mathbf{x_N^i}) \in D} \alpha_i(\kappa(\mathbf{x}, \mathbf{x_P^i}) - \kappa(\mathbf{x}, \mathbf{x_N^i})) \tag{17}$$

where α_i are non-negative coefficients, $\kappa(\mathbf{x_j x_j})$ is the kernel function and $(\mathbf{x_P^i}, \mathbf{x_N^i})$ are pairs of training samples.

For all experiments reported in this paper, we use a Gaussian projection (i.e. Gaussian kernel) as it can generate a wide range of non-linear functions while it contains one single parameter (γ). We set $C = 0.001$ and $\gamma = 1.0$ after systematic adjustment of their values.

5.3 Automatic Feature Selection

Feature selection is a procedure commonly used in data mining to reduce the dimensionality of training data by removing features that seemingly do not contain relevant information about the function modelled. The basic procedure consists of evaluating several combinations of features using a predefined *fitness function*. In this paper we use a genetic algorithm to search the space of all possible combinations of features — this is known as Genetic-search Feature Selection (GFS) [40]. We use a population of 19 bit-chromosomes — their length being the total number of features extracted — that represent whether a particular

feature is selected (1) or not (0). Across 15 iterations, pairs of feature subsets are selected based on a ranking selection method (the higher the fitness of a feature subset, the greater the probability of being selected) and recombined via uniform crossover (probability of 0.8) to generate new feature subsets (offspring). A mutation operator adds or removes one feature to the offspring's chromosome with probability 0.01.

The fitness of each subset of features is calculated as the average 10-fold cross-validation (CV) accuracy of an SVM employing a Gaussian kernel trained on the available data using only the selected subset of features.

6 Results and Analysis

To measure the degree of agreement between the values provided by SD, NH, GC, FI and TGB and the crowdsourced self-reports, we calculate the correlation coefficients between them, following the statistical analysis procedure for pairwise preference data presented in [57] using the test statistic:

$$c(\mathbf{z}) = \sum_{i=1}^{N} \{z_i/N\} \tag{18}$$

where N is the total number of samples to correlate, $z_i = +1$, if the video reported as more fair in pair i yields a higher metric than the video reported as less fair, and $z_i = -1$, if the video reported as more fair in pair i yields a lower metric than the video reported as less fair. In the calculation of $c(\mathbf{z})$ we only take into account clear preferences, that is, we only consider pairs in which a clear preference — A is more fair than B or B is more fair than A (2-AFC) — is expressed, that is, $N = 147$. The p-values of $c(\mathbf{z})$ are obtained via the binomial distribution. Tables 3(a), 3(b), 3(c), 3(d) and 3 present the $c(\mathbf{z})$ values and their corresponding p-values, for each scenario and in total, for the metrics SD, NH, CG, FI and TGB respectively. The first three columns (after the "Scenario" column) report the number of choices for the alternatives of the 4-AFC. Columns "Match" and "Mismatch" represent the number of 2-AFC preferences that respectively, match and mismatch the metric value.

6.1 Validating the Ad-hoc Metrics of Fairness

The first observable result is that all four metrics (SD, NH, GC and FI — see Tables 3(a), 3(b), 3(c) and 3(d)) appear to be consistent with the reported preferences. NH — and consequently FI, which presents the exact same results — yield correlation values above 0.7 for the 4 receivers scenario, whilst the $c(\mathbf{z})$ values are not as high (just above 0.4) for the 2, 3, and 5 receiver scenarios. A similar behaviour and $c(\mathbf{z})$ values are observed for the GC metric.

Even though SD appears to be consistent with the notion of fairness, not surprisingly, it is the metric which scores the lowest correlation coefficients, since

Table 3. Analysis of correlation $c(\mathbf{z})$ between fairness metrics and reported fairness. Significant $c(\mathbf{z})$ values appear in bold — significance is 5%.

(a) SD

Scenario	2-AFC Preference	Equally Fair	Both Unfair	Match	Mismatch	$c(\mathbf{z})$	p-value
2 receivers	38	9	13	28	10	**0.47**	0.03
3 receivers	41	4	15	24	17	0.17	0.45
4 receivers	36	11	13	31	5	**0.72**	< 0.01
5 receivers	32	6	22	24	8	**0.50**	0.01
Total	147	30	63	107	40	**0.46**	0.02

(b) NH

Scenario	2-AFC Preference	Equally Fair	Both Unfair	Match	Mismatch	$c(\mathbf{z})$	p-value
2 receivers	38	9	13	29	9	**0.53**	0.01
3 receivers	41	4	15	29	12	**0.41**	0.05
4 receivers	36	11	13	31	5	**0.72**	< 0.01
5 receivers	32	6	22	25	7	**0.56**	< 0.01
Total	147	30	63	114	33	**0.55**	< 0.01

(c) GC

Scenario	2-AFC Preference	Equally Fair	Both Unfair	Match	Mismatch	$c(\mathbf{z})$	p-value
2 receivers	38	9	13	29	9	**0.53**	0.01
3 receivers	41	4	15	29	12	**0.42**	0.05
4 receivers	36	11	13	30	6	**0.67**	< 0.01
5 receivers	32	6	22	25	7	**0.56**	< 0.01
Total	147	30	63	113	34	**0.54**	< 0.01

(d) FI

Scenario	2-AFC Preference	Equally Fair	Both Unfair	Match	Mismatch	$c(\mathbf{z})$	p-value
2 receivers	38	9	13	29	9	**0.53**	0.01
3 receivers	41	4	15	29	12	**0.41**	0.05
4 receivers	36	11	13	31	5	**0.72**	< 0.01
5 receivers	32	6	22	25	7	**0.56**	< 0.01
Total	147	30	63	114	33	**0.55**	< 0.01

(e) TGB

Scenario	2-AFC Preference	Equally Fair	Both Unfair	Match	Mismatch	$c(\mathbf{z})$	p-value
2 receivers	38	9	13	27	11	**0.42**	0.06
3 receivers	41	4	15	35	6	**0.71**	< 0.01
4 receivers	36	11	13	31	5	**0.72**	< 0.01
5 receivers	32	6	22	23	9	**0.44**	0.04
Total	147	30	63	116	31	**0.58**	0.01

standard deviation is a measure of dispersion of the data with respect to a reference mean value. Although SD correlates well with the human notion of fairness when highly fair and unfair strategies are existent, this does not necessarily hold for strategies of intermediate levels of fairness.

Overall, TGB (see Table 3) scores a correlation coefficient higher (but not significantly higher) than any other ad-hoc metric. Therefore, we can state that TGB, NH — and consequently GC and FI — represent the notion of fairness for the one-to-many interaction scenario equally well overall. TGB yields correlation values above 0.7 for the 3 and 4 receivers scenario, whilst the $c(\mathbf{z})$ values are not as high (just above 0.4) for the 2 and the 5 receivers scenarios.

Compared to NH and FI, TGB manages to improve the correlation with the perceived fairness for the 3 receivers scenario, however, at a cost of more mismatches for the 2 and the 5 receivers scenarios. The 2 and 4 receivers scenarios are those for which the $c(\mathbf{z})$ values of TGB, NH, GC and FI are almost identical. There is no doubt that 3 receivers is the most complex scenario among the 4 we have considered in our experimental setup: there is a difference in the group sizes ($n_{\text{red}} = 2$ and $n_{\text{blue}} = 1$), a high amount of resources ($m = 4$), a big difference in satisfaction within the red group itself, and finally, the resource values are generally low, except for one, for which $v_i = 0.9$ — see Table 1. The 3 receiver scenario appears to have instigated different and complicated perceptions of fairness among the participants, which can be better captured by TGB, as opposed to the other metrics.

NH, GC and FI outperform TGB in the 5 receivers scenario as the metrics generate two very different orderings of the strategies. It appears that the large population size and the many available resources lead to a lower impact of existing group structures on the perception of fairness. In support of this hypothesis many submitted comments of the participants suggest that a fair distribution strategy should first prioritise the fulfilment of all receivers' satisfaction independently of their group identities and, only subsequently, distribute the resources according to a group-based strategy.

The crowdsourced reports highlight that it is easier to report a clear fairness preference (i.e. 2-AFC) for scenarios with a lower population (2 and 3 receivers) rather than scenarios with a high population (4 and 5 receivers). This finding suggests that there might be potential difficulties in observing and distinguishing complex distribution strategies within our 3D game design. It is worth noting in this respect that we received only three additional comments related to the difficulty to perceive the distribution strategies; however, these were submitted by inexperienced players, who spent only up to two hours on gaming per week, as we could retrieve from their demographic entries.

6.2 Modelling Fairness via Preference Learning

As an alternative to ad-hoc metric design, we investigated the inverse approach and followed a data-driven methodology to construct a model that is directly built on the crowdsourced pairwise preferences, to be compared to the ad-hoc metrics. For that purpose, we run GFS 10 times and pick the feature subset that

Table 4. Average and best performance across 10 trials of the *rank SVM* algorithm. Performance accuracy is assessed through 10-fold cross-validation. Correlation values ($c(\mathbf{z})$) are derived from the 10-fold CV accuracy.

Feature Set		Accuracy (%)	Matches	Mismatches	$c(\mathbf{z})$
Random Features	Average	73.54	108	39	0.47
	Best	77.09	113	34	0.54
All Features	Average	74.95	110	37	0.50
	Best	76.81	112	35	0.52
Best Feature Subset	Average	79.33	117	30	0.59
	Best	81.86	120	27	0.63

feeds an SVM model (as described in Section 5) which yields the highest 10-fold CV accuracy on the pairwise preference data. In order to reduce the impact of non-determinism existent in the separation of the data into folds, we run 10 trials of the algorithm using three different feature sets: the best-performing feature set, the set that contains all 19 features extracted, and randomly selected features. Table 4 reports the average and highest accuracies and the corresponding $c(\mathbf{z})$ values of the three different feature sets. The best performing feature set yields accuracies which are significantly higher (tested via a t-test) when compared to the full feature set (p-value < 0.01) and the randomly-selected feature subset (p-value < 0.01). Thus, it appears that genetic feature selection (GFS) improves the accuracy of the model, on average, (79.33%) as it outperforms randomly selected features (73.54%) and all features (74.95%) considered.

The best-feature set ($c(\mathbf{z}) = 0.63$) supports a model that outperforms the correlation coefficient of the TGB metric. This model predicts 81.86% of the pairwise fairness reports correctly relying on five features selected by GFS: the initial average satisfaction value for the blue group (μ^0_{blue}); the average satisfaction of the red group after the second delivery (μ^2_{red}); the average satisfaction of the blue group after the second and third delivery (μ^2_{blue} and μ^3_{blue}, respectively); and the satisfaction provided at the third delivery (v^3_j). The selected feature subset suggests that particular resource deliveries to particular groups are of key importance for determining and approximating fairness.

7 Discussion

The TGB, NH, GC and FI ad-hoc metrics manage to represent the notion of fairness well, as the cross-validation analysis performed with the data gathered from the crowdsourcing experiment showed high consistency and strong statistical significance. The question of how to quantitatively model fairness precisely, in order to subsequently infer the presence of social preferences, collaboration and global patterns such as group identities has been answered, though only in part. The key findings of the paper can evidently contribute to further investigations for addressing the aforementioned question.

The difference in performance between TGB and the NH, GC and FI metrics suggests that the context-based, expert-driven metric TGB, might have introduced some bias over-fitting to the examined scenario. It seems therefore intuitive that, prior to considering TGB as a universal metric for one-to-many interaction scenarios, similarly to NH, GC and FI, more studies should be conducted. For instance, with respect to the scenario's formal definition (Eq. 1), studies based on a higher number of agents, groups, resources, and resource allocations, could be made. On the other hand, NH, GC and FI showed good efficacy with respect to modelling fairness, even in the one-to-many scenario. Finally, SD showed a low correlation with the perceived notion of fairness; this is explained by the nature of the metric, which describes the dispersion of the data based on a reference mean value. This suggests that fairness is an absolute notion, rather than being relative to a reference satisfaction value. As a consequence, it is likely that similar measures of fairness, e.g. the coefficient of variation [29], would show similar performances.

By following our assumption that fairness is a feature of interactions which can help with the identification of preferred individuals in the population, hence group structures, we have been investigating methods to use fairness as a feature for collaboration learning in order to detect the formation and consolidation of group identities in complex artificial societies of believable, human-like artificial agents [23, 24]. These agents manifested *reciprocal* and *altruistic* social preferences, interacting with each other, iteratively, by means of the ultimatum game. The interaction scenario was interpreted as a sequence of one-to-many interactions between one provider agent and many receiver agents, and NH was used to calculate the fairness of the providers' offers. The results obtained showed that NH can help with the detection of existing group structures and is robust across different population sizes, group structure typologies, and in presence of diverse locality of interactions among the agents.

Future work would intuitively focus on the investigation of other metrics of fairness. We hereby suggest either the definition of new metrics, or the identification of existing ones which would put an emphasis on the sequence of the resources being distributed, which is only partly achieved by TGB. Moreover, fairness can also be associated with a number of other complex notions, such as balance [38] and asymmetry [41]. Finally, due to the positive correlation coefficients scored by the Fairness Index (FI), future studies will aim to investigate how well other FI-related metrics [43] are linked to our scenario and how well they would correlate to the crowdsourced self-reports.

While the SVM approach yielded high-performing fairness models (model accuracy $> 80\%$) that surpass the correlation of ad-hoc metrics with the crowdsourced data, the generalisability of the model to other settings is likely to be lower as it is built on data and features from a particular environment. Nevertheless, as the accuracy is evaluated on data unseen during training, it is expected to maintain its superiority within similar settings. Furthermore, the expressivity of the metrics over the black-box Gaussian SVM model provides a key advantage for their use. On that basis, more preference learning algorithms will

be tested and compared: possible candidates for learning the mapping between pairwise preferences and social dynamics in the game include bayesian [8] and neuro-evolutionary preference learning [39]. Towards the data-driven approach to modelling fairness, more experimental data will be required from diverse and dissimilar game scenarios containing variant numbers of agents, groups of agents and initial conditions.

The self-reports and some of the extra comments filled by the participants suggest that, particularly for occasional and non-gamers, the 3D game-based implementation of the one-to-many resource allocation scenario, with an emphasis on how to represent the levels of satisfaction of the agents, might add a bottleneck with respect to the perception of fairness. This could also be the reason why the participants are more *confident* to report clear preferences (A is more fair than B or B is more fair than A) in scenarios with smaller populations. This drawback could be reduced by allowing the participants to have a more active role, rather than just following the provider agent and observing its allocation strategy. For instance, future work could be focused on letting the participants play the role of the provider agent, distribute resources, and subsequently describe the strategies they adopted. Although this approach might introduce challenges on the quantification of the strategy descriptions, it could on the other hand allow for the discovery of alternative, highly complex distribution strategies.

The introduction of group identities in the population was motivated by the intention to represent the existence of social preferences under the perspective of the provider agent. The differences of group identities are to be found, solely, on the colour of the body of the receiver agents. Moreover, the provider agent, as depicted in both Figure 1 and Figure 2, does not explicitly belong neither to the red nor to the blue group. Although there is a vast corpus of studies suggesting that group behaviours and identities can be observed independently on how arbitrarily the groups are instantiated (see [1, 5, 6, 9] among others), the game design we adopted might not represent real-life, global structures, e.g. ethnicity or friendship well and might explain why NH, GC and FI — which are group-independent — correlate well with the self-reported data. Further work on the enrichment of the graphical representation of the group identities would be considered.

Although some of the motivations which led us to consider only two groups were driven by the need to represent the dichotomy *in-group* vs. *out-group* [9] (see Section 3.2), it might be possible that some of the metrics, especially TGB, might not be able to scale well in presence of more complex scenarios. On the other hand, more complex scenarios would lead to the increase of the number of features describing the distribution strategies. As a consequence, it might be possible that those features extracted via GFS for rank SVM would become more generic, as opposed to those used by the best performing rank SVM for the current two-group scenario (see Section 6.2). Given that we cannot clearly foresee the changes in the consistency of NH, GC and FI — since more complex scenarios would lead to a wider plethora of group-based distribution strategies,

and hence their human perception — future work based on scenarios with more than two will be considered.

The proposed crowdsourcing approach for metric design proves that it is possible to design accurate measures of fairness. Beyond our resource allocation scenario, the proposed crowdsourcing approach can be used as a validation tool to explain the discrepancies between the results obtained in evolutionary games (i.e. based on artificial societies) and those found in nature [25]. Preliminary results, based on artificial societies [23, 24], suggest that fairness is a feature of interactions which would expose the preference of individuals. Thus, the metrics TGB, NH, GC and FI (or any other metrics which correlate well to those) can be used to detect unfair treatments which may lead to social conflicts [7, 61]. Similarly, fairness metrics can be used to extract student profiles in collaborative educational games [21, 61].

8 Conclusions

This paper addressed the problem of quantitatively measuring fairness, under the perspective of one individual who interacts with multiple other individuals (i.e. *one-to-many* interaction scenario). Given that fairness is an abstract and ambiguous term with fuzzy boundaries, we have relied on crowdsourced data obtained via pairwise preference self-reports, and used it to cross validate six metrics of fairness. The first four metrics are well-established metrics of the dispersion of data, namely standard deviation, normalised entropy, the Gini coefficient, and the Fairness Index. The fifth metric, called temporal group-based fairness (TGB), is a new metric proposed by the authors, is ad-hoc designed for the one-to-many interaction scenario, and takes into consideration context-based aspects of the distribution task, such as the sequence of distribution and the presence of group structures within the receiver agents. Finally, the sixth metric is machine learned on the preference data via ranking Support Vector Machines (SVM).

The results obtained show that all metrics are highly consistent (though with different degrees of consistency) with the perception of fairness of hundreds of our survey participants. It seems, however, that the temporal group-based metric is expressive enough and captures fairness more accurately than the other ad-hoc metrics. Even though the SVM model yields the most accurate fairness measure, the TGB metric is far more expressive and usable. The normalised entropy and fairness index metrics, however, appear to be the most appropriate for context-free and generic use, as the TGB metric is based on (and tightened to) the context of the one-to-many resource allocations scenario. Preliminary results have shown the efficacy of NH in capturing fairness and collaboration in artificial societies of agents that play the social ultimatum game [23, 24].

The fairness metrics proposed can be used in both simulated scenarios of artificial agent societies to investigate global phenomena, such as collaboration and the emergence of group structures [23, 24], or in educational collaborative virtual

environments, in which human-controlled avatars interact with each other [21, 61].

Acknowledgments. This work has been supported, in part, by the FP7 ICT projects SIREN (project no: 258453) and ILearnRW (project no: 318803). The authors would like to thank all participants of the crowdsourcing experiment. Special thanks to Yana Knight for proofreading.

References

1. Akerlof, G.A., Kranton, R.E.: Economics and Identity. The Quarterly Journal of Economics 115(3), 715–753 (2000)
2. Axelrod, R., Hamilton, W.D.: The Evolution of Cooperation (1981)
3. Bolton, G.E., Katok, E., Zwick, R.: Dictator Game Giving: Rules of Fairness Versus Acts of Kindness. International Journal of Game Theory 27, 269–299 (1998)
4. Charness, G., Rabin, M.: Understanding Social Preferences with Simple Tests. The Quarterly Journal of Economics 117(3), 817–869 (2002)
5. Charness, G., Rigotti, L., Rustichini, A.: Individual Behavior and Group Membership. Available at SSRN 894685 (2006)
6. Chen, Y., Li, S.X.: Group Identity and Social Preferences. The American Economic Review pp. 431–457 (2009)
7. Cheong, Y.G., Khaled, R., Grappiolo, C., Campos, J., Martinho, C., Ingram, G.P.D., Paiva, A., Yannakakis, G.N.: A Computational Approach Towards Conflict Resolution for Serious Games. In: Proceedings of the Sixth International Conference on the Foundations of Digital Games. ACM (2010)
8. Chu, W., Ghahramani, Z.: Preference Learning with Gaussian Processes. In: Proceedings of the 22nd International Conference on Machine learning. ACM (2005)
9. Dawes, R.M., Messick, D.M.: Social Dilemmas. International Journal of Psychology 2(35), 111–116 (2000)
10. De Jong, S., Tuyls, K., Verbeeck, K.: Artificial Agents Learning Human Fairness. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems. pp. 863–870 (2008)
11. Dianati, M., Shen, X., Naik, S.: A New Fairness Index for Radio Resource Allocation in Wireless Networks. In: Wireless Communications and Networking Conference. vol. 2, pp. 712–717 (2005)
12. Ducheneaut, N., Yee, N., Nickell, E., Moore, R.J.: Alone Together? Exploring the Social Dynamics of Massively Multiplayer Online Games. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 407–416. ACM (2006)
13. Eagle, N., Pentland, A.S., Lazer, D.: Inferring Friendship Network Structure by Using Mobile Phone Data. Proceedings of the National Academy of Sciences 106(36), 15274–15278 (2009)
14. Epstein, J.M., Axtell, R.L.: Growing Artificial Societies: Social Science from the Bottom Up (Complex Adaptive Systems). The MIT Press (1996)
15. Fehr, E., Fischbacher, U.: Why Social Preferences Matter — The Impact of Non-Selfish Motives on Competition, Cooperation and Incentives. Economic Journal 112, 1–33 (2002)

16. Fehr, E., Schmidt, K.M.: A Theory Of Fairness, Competition, And Cooperation. The Quarterly Journal of Economics 114(3), 817–868 (August 1999)
17. Forsythe, R.: Fairness in Simple Bargaining Experiments. Games and Economic Behavior 6(3), 347–369 (1994)
18. Gal, Y., Grosz, B.J., Kraus, S., Pfeffer, A., Shieber, S.: Colored Trails: a Formalism for Investigating Decision-making in Strategic Environments. In: Proceedings of the 2005 IJCAI Workshop on Reasoning, Representation, and Learning in Computer Games. pp. 25–30 (2005)
19. Gini, C.: Measurement of Inequality of Incomes. The Economic Journal 31(121), 124–126 (March 1921)
20. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99(12), 7821–7826 (2002)
21. Grappiolo, C., Cheong, Y.G., Khaled, R., Yannakakis, G.N.: Modelling Global Pattern Formation for Collaborative Learning Environments. In: Proceedings of the IEEE International Conference on Advanced Learning Technologies (2012)
22. Grappiolo, C., Cheong, Y.G., Togelius, J., Khaled, R., Yannakakis, G.N.: Towards Player Adaptivity in a Serious Game for Conflict Resolution. In: Proceedings of the 3rd IEEE International Conference in Games and Virtual Worlds for Serious Applications. pp. 192–198 (2011)
23. Grappiolo, C., Togelius, J., Yannakakis, G.N.: Interaction-based Group Identity Detection via Reinforcement Learning and Artificial Evolution. In: Proceedings of the Evolutionary Computation and Multi-agent Systems and Simulation workshop, Genetic and Evolutionary Computation Conference. pp. 1423–1430. ACM (2013)
24. Grappiolo, C., Yannakakis, G.N.: Towards Detecting Group Identities in Complex Artificial Societies. In: Proceedings of the Simulation of Adaptive Behaviour Conference. pp. 421–430 (2012)
25. Greenwood, G.W., Ashlock, D.: Evolutionary Games and the Study of Cooperation: Why Has So Little Progress Been Made? In: Proceedings of the IEEE World Congress on Computational Intelligence (2012)
26. Hammond, R.A., Axelrod, R.: The Evolution of Ethnocentrism. Journal of Conflict Resolution 50(6), 926–936 (2006)
27. Herbrich, R., Graepel, T., Obermayer, K.: Support Vector Learning for Ordinal Regression. In: Proceedings of the International Conference on Artificial Neural Networks. vol. 1, pp. 97 –102 (1999)
28. Huberman, B.A., Glance, N.S.: Evolutionary Games and Computer Simulations. Proceedings National Academy of Science 90(16), 7716–7718 (1993)
29. Jain, R., Chiu, D.M., Hawe, W.R.: A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer System. Eastern Research Laboratory, Digital Equipment Corporation (1984)
30. Joachims, T.: Learning to Classify Text Using Support Vector Machines — Methods, Theory, and Algorithms. Kluwer/Springer (2002)
31. Joachims, T.: Optimizing Search Engines Using Clickthrough Data. In: Proceedings of the 8th SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 133–142. ACM (2002)
32. Joe-Wong, C., Sen, S., Lan, T., Chiang, M.: Multi-resource Allocation: Fairness-efficiency Tradeoffs in a Unifying Framework. In: Proceedings of the IEEE International Conference on Computer Communications. pp. 1206–1214. IEEE (2012)
33. Kagel, J.H., Kim, C., Moser, D.: Fairness in Ultimatum Games with Asymmetric Information and Asymmetric Payoffs. Games and Economic Behavior 13(1), 100–110 (1996)

34. Kim, J.H.: The Role of Identity in Intra-and Inter-Group Bargaining in the Ultimatum Game. Undergraduate Economic Review 4(1), 6 (2008)
35. Kranton, R., Pease, M., Sanders, S., Huettel, S.: Identity, Group Conflict, and Social Preferences. Working Paper (2012)
36. Lancichinetti, A., Fortunato, S.: Limits of Modularity Maximization in Community Detection. Physical Review E 84(6), 066122 (2011)
37. Lansing, S.J.: Complex Adaptive Systems. Annual Review of Anthropology 32, 183–204 (2003)
38. Mahlmann, T., Togelius, J., Yannakakis, G.N.: Modelling and Evaluation of Complex Scenarios with the Strategy Game Description Language. In: Proceedings of the IEEE Conference for Computational Intelligence and Games. Seoul, KR (2011)
39. Martínez, H.P., Yannakakis, G.N.: Mining multimodal sequential patterns: a case study on affect detection. In: Proceedings of International Conference on Multimodal Interfaces (ICMI). pp. 3–10. ACM (2011)
40. Martínez, H., Yannakakis, G.: Genetic Search Feature Selection for Affective Modeling: a Case Study on Reported Preferences. In: Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments. pp. 15–20. ACM (2010)
41. Martinez, R., Kay, J., Wallace, J., Yacef, K.: Modelling Symmetry of Activity as an Indicator of Collocated Group Collaboration. In: User Modeling, Adaption and Personalization, vol. 6787, pp. 207–218. Springer Berlin / Heidelberg (2011)
42. Marzo, F., Grosz, B.J., Pfeffer, A.: Social preferences in Relational Contexts. In: In Fourth Conference in Collective Intentionality (2005)
43. Montuno, K., Zhacfi, Y.: Fairness of Resource Allocation in Cellular Networks: A Survey. Resource Allocation in Next Generation Wireless Networks pp. 249–266 (2006)
44. Nowak, M.A.: Five Rules for the Evolution of Cooperation. Science 314(5805), 1560–1563 (2006)
45. Palla, G., Barabási, A.L., Vicsek, T.: Quantifying Social Group Evolution. Nature 446(7136), 664–667 (2007)
46. Pandremmenou, K., Kondi, L.P., Parsopoulos, K.E.: Fairness Issues in Resource Allocation Schemes for Wireless Visual Sensor Networks. In: IS&T/SPIE Electronic Imaging. pp. 866601–866601. International Society for Optics and Photonics (2013)
47. Prada, R., Paiva, A.: Teaming Up Humans with Autonomous synthetic Characters. Artificial Intelligence 173(1), 80–103 (2009)
48. Rabin, M.: Incorporating Fairness into Game Theory and Economics. The American Economic Review pp. 1281–1302 (1993)
49. Rocha, J.B., Mascarenhas, S., Prada, R.: Game Mechanics for Cooperative Games, pp. 73–80. Universidade do Minho (2008)
50. Seif El-Nasr, M., Aghabeigi, B., Milam, D., Erfani, M., Lameman, B., Maygoli, H., Mah, S.: Understanding and Evaluating Cooperative Games. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems. pp. 253–262. ACM (2010)
51. Shaker, N., Yannakakis, G., Togelius, J.: Crowd-Sourcing the Aesthetics of Platform Games. IEEE Transactions on Computational Intelligence and AI in Games (2012)
52. Shi, H., Prasad, R.V., Rao, V.S., Niemegeers, I.: A Fairness Model for Resource Allocation in Wireless Networks. In: Networking Workshops. pp. 1–9. Springer (2012)
53. Sonntagbauer, P., Aizstrauts, A., Ginters, E., Aizstrauta, D.: Policy Simulation and E-Governance. In: IADIS International Conference e-Society (2012)

54. Szell, M., Thurner, S.: Measuring Social Dynamics in a Massive Multiplayer Online Game. Social Networks 32(4), 313–329 (2010)
55. Tan, G., Guttag, J.V.: Time-based Fairness Improves Performance in Multi-Rate WLANs. In: USENIX Annual Technical Conference, General Track. pp. 269–282 (2004)
56. Xianyu, B.: Social Preference, Incomplete Information, and the Evolution of Ultimatum Game in the Small World Networks: An Agent-Based Approach. Journal of Artificial Societies and Social Simulation 13(2), 7 (2010)
57. Yannakakis, G.N., Hallam, J.: Towards Optimizing Entertainment in Computer Games. Applied Artificial Intelligence 21(10), 933–971 (2007)
58. Yannakakis, G.N., Hallam, J.: Rating vs. Preference: A Comparative Study of Self-reporting. In: Springer (ed.) Proceedings of the 2011 Affective Computing and Intelligent Interaction Conference. Springer (2011)
59. Yannakakis, G.N., Martnez, H.P., Jhala, A.: Towards Affective Camera Control in Games. User Modeling and User-Adapted Interaction 20, 313–340 (2010)
60. Yannakakis, G.N., Togelius, J.: Experience-Driven Procedural Content Generation. IEEE Transactions on Affective Computing 2, 147–161 (2011)
61. Yannakakis, G.N., Togelius, J., Khaled, R., Jhala, A., Karpouzis, K., Paiva, A., Vasalou, A.: Siren: Towards Adaptive Serious Games for Teaching Conflict Resolution. In: Proceedings European Conference on Games-Based Learning (ECGBL). pp. 412–417. Copenhagen (2010)