Multiplayer Ultimatum Game in Populations of Autonomous Agents

Fernando P. Santos,
Francisco C. Santos
INESC-ID and Instituto Superior Técnico,
Universidade de Lisboa
Taguspark, Av. Prof. Cavaco Silva
2780-990 Porto Salvo, Portugal
{fernando.pedro,franciscocsantos}@ist.utl.pt

Francisco S. Melo,
Ana Paiva
INESC-ID and Instituto Superior Técnico,
Universidade de Lisboa
Taguspark, Av. Prof. Cavaco Silva
2780-990 Porto Salvo, Portugal
{fmelo,ana.paiva}@inesc-id.pt

Jorge M. Pacheco
CBMA and Departamento de Matemática e
Aplicações, Universidade do Minho
Campus de Gualtar
4710-057 Braga, Portugal
jmpacheco@math.uminho.pt

ABSTRACT

There are numerous human decisions and social preferences whose features are not easy to grasp mathematically. Fairness is certainly one of the most pressing. In this paper, we study a multiplayer extension of the well-known Ultimatum Game through the lens of a reinforcement learning algorithm. This game allows us to study fair behaviors beyond the traditional pairwise interaction models. Here, a proposal is made to a quorum of Responders, and the overall acceptance depends on reaching a threshold of individual acceptances. We show that, while considerations regarding the sub-game perfect equilibrium of the game remain untouched, learning agents coordinate their behavior into different strategies, depending on factors such as the group acceptance threshold, the group size or disagreement costs. Overall, our simulations show that stringent group criteria trigger fairer proposals and the effect of group size on fairness depends on the same group acceptance criteria. Fairness can be boosted by the imposition of disagreement costs on the Proposer side.

1. INTRODUCTION

The role of fairness in decision-making has for long captured the attention of academics and the subject comprises a fertile ground of multidisciplinary research [9, 10]. In this context, the Ultimatum Game (UG), proposed more than thirty years ago, stands as a simple interaction paradigm that is capable of capturing the essential clash between rationality and fairness [12]. In its original form, two players interact acquiring two distinct roles: Proposer and Responder. The Proposer is endowed with some resource and has to propose a division to the second player. After that, the Responder has to state her acceptance or rejection. If the proposal is rejected, none of the players earns anything. If the proposal is accepted, they will divide the resource as it was proposed. A fair outcome is usually defined as an egalitarian division, in which both the Proposer and the Responder earn a similar reward.

The minimalism of this game is convenient to allow a mathematical treatment that aims at computing the most probable outcome in which humans will end up, while playing it. A first approach would be to look into each agent as being rational and oriented to the maximization of rewards. Thinking in a backward fashion, one may realize that the Responder should always accept any offer; the Proposer, confident about this reasonable reaction, should always propose to give the minimum possible amount to the Responder. Indeed, this line of thought gives an intuition for the sub-game perfect equilibrium of the UG: low offers by Proposers and low acceptance thresholds by Responders [19]. These predictions regarding how people act are, however, misleading. A vast number of works report experiments with people in which they behave very differently from the rational sub-game prediction [23, 12, 30]. Humans tend to reject low proposals, i.e., they have high thresholds of acceptance and they tend to offer fair divisions. The explanations for this fact diverge. Some authors argue that the Proposers have a natural propensity to be fair; others suggest that they fear to have a proposal rejected [30]. Interestingly, humans keep exhibiting fair preferences in dictator games, where a proposal is always accepted no matter what is the Responder opinion [10], a behavior that can be explained by reciprocity-like mechanisms [13].

The mathematical treatment of this game followed the need to come up with different predictions, other than the game theoretical sub-game perfect equilibrium. Why is that it pays for individuals to reject low proposals and offer high ones? How to explain the evolution of this behavior mathematically? Resorting to evolutionary game theoretical tools, Nowak et. al. suggested that if Proposers are able to get pieces of information about previous actions of the opponents, then it is worth for the Responders to cultivate a fierce reputation [18]. This way, Proposers would offer more to Responders that are used to reject low offers and it naturally leads the Responders to nurture an intransigent reputation by rejecting unfair offers. Other models attribute the evolution of fairness to the repetition of interactions [33] or empathy [21]. A slightly different approach suggests that

fair Proposers and Responders may emerge due to the topological arrangement of their network of contacts: if individuals are arranged in lattices [22, 28] clusters of fairness may emerge. Also using learning frameworks, a lot of attention was given to the UG [11, 6, 5]. For a neat work that combines learning agents (that play UG) with complex networks, volunteering and reputation we refer to De Jong et al. [6].

Any mathematical explanation (and/or prediction) for human behavior in the UG holds as a fundamental result of clear importance in areas as evolutionary biology, economics or philosophy. In Artificial Intelligence specifically, these advancements provide an important asset for the design of artificial agents and the simulation of artificial societies, in terms of *i) performance*, *ii) expectation* and *iii) accuracy*: *i)* artificial agents that do incorporate features of human-like behavior when playing the UG are agents capable of performing better (a purely selfish agent that always offers close to nothing to a human Responder will naturally be doomed to a hopeless performance) [5, 15]; *ii)* artificial agents playing with humans in UG-like interactions are naturally more believable and enjoyable if they exhibit human preferences as they will meet their opponents expectation; *iii)* models based on the simulation of artificial societies that seek to predict the impact of policies on aggregate behavior and emergent outcomes [29], will be more accurate if they include the appropriate mathematical assumptions regarding human behaviors; in this case, the proper feelings towards fairness and unfairness.

While these stand as important criteria for the case of agents playing the two-player UG, the same apply to a wide range of human-agent interactions that a pairwise interaction model does not enclose. It is perfectly straightforward to realize that also UG instances take place in groups, with proposals being made to assemblies [25]. Take the case of pervasive democratic institutions, economic and climate summits, collective bargaining, markets, auctions, or the ancestral activities of proposing divisions regarding the loot of group hunts and fisheries. All those examples go beyond a pairwise interaction. Indeed, there is a growing interest in doing experiments with multiplayer versions of the UG [11, 9, 16, 7]. A simple extension may turn it adequate to study a wide variety of ubiquitous formats of people encounters. This extension, the Multiplayer UG (MUG), allows to study the traditional 2-person UG in a context where proposals are made to groups and the groups should decide, through suffrage, about its acceptance or rejection.

In the context of this game, if we want to fulfill the previous criteria, some immediate answers need to be addressed: what is the role of group in the individual decisions? What is the impact of group acceptance rules on individual offers? What is the role of group size on fairness?

In this paper we provide a model to approach those answers, by combining MUG with agents that learn how to play it through reinforcement learning [27]. We test the well-known Roth-Erev algorithm [23]. We show that there is a set of parameters (group size, decision rule, disagreement costs) that are relevant given the setting of MUG and that provide non-trivial effects regarding the learned strategies.

We start by reviewing the equilibrium notions of classical game theory, namely, the sub-game perfection. We show that the above game parameters are irrelevant regarding the equilibrium approach. Notwithstanding, they deeply affect the learned behaviors, with serious impacts on group fairness.

Table 1: Glossary

Symbol	Meaning
p	Offer by Proposer
q	Acceptance threshold of Responder
$\Pi_P(p_i, q_{-i})$	Payoff earned by a Proposer
$\Pi_R(p_j, q_{-j})$	Payoff earned by a Responder
$\Pi(p_i, q_i, p_{-i}, q_{-i})$	Payoff being Proposer and Responder
$a_{p_i,q_{-i}}$	Group acceptance flag
d	Disagreement cost
$Q(t)$	Propensity matrix at time t
λ	Forgetting rate
ϵ	Local experimentation
$\rho_{ki}(t)$	Probability that k uses strategy i
$\bar{p}$	Average p population-wide
$i_{p,q}$	Integer representation of strategy (p, q)
R	Number of runs
Z	Population size
N	Group size
M	Group acceptation threshold
T	Number of time steps
R	Number of runs

2. MULTIPLAYER ULTIMATUM GAME

In the typical pairwise UG, a Proposer receives a sum and decides the fraction (p) that should offer to a Responder. The Responder must then state her acceptance or rejection. This decision can rely on a personal threshold (q), which is used to decide about acceptance or rejection: if $p \geq q$ the proposal is accepted and if $p < q$, the proposal is rejected. Considering that the amount being divided sums to 1, if the proposal is accepted the Proposer earns $1 - p$ and the Responder earns p. If the proposal is rejected, none of the individuals earn anything [18].

This two-person game can now be extended to an N-person game, assuming the existence of a quorum of $N - 1$ Responders. Again, a proposal is made (p), yet now each of the Responders states acceptance of rejection and the overall acceptance depends on an aggregation of these individual decisions: if the number of acceptances equals or exceeds a threshold M, the proposal is accepted by the group. In this case, the Proposer keeps what she did not offer $(1 - p)$ and the offer is evenly divided by all the Responders $(p/(N-1))$; otherwise, if the number of acceptances remains below M, the proposal is rejected by the group an no one earns anything.

The payoff function describing the gains of a Proposer i, with strategy p_i, facing a quorum of Responders with strategies $q_{-i} = \{q_1, ..., q_j, ..., q_{N-1}\}, j \neq i$ reads as

$$\Pi_P(p_i, q_{-i}) = (1 - p_i)a_{p_i,q_{-i}} \tag{1}$$

Where $a_{p_i,q_{-i}}$ summarises group acceptance of the proposal made by agent i, p_i, standing as

$$a_{p_i,q_{-i}} = \begin{cases} 1, & \text{if } \sum_{q_j \in q_{-i}} \Theta(p_i - q_j) \geq M. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

$\Theta(x)$ is the Heaviside unit step function, having value 1

whenever $x \geq 0$ and 0 otherwise. This way, $\Theta(p_i - q_j) = 1$ if agent j accepts agent's i proposal.
Similarly, the payoff function describing the gains of a Responder belonging to a quorum with a strategy profile $q_{-j} = \{q_1, ..., q_k, q_i, ..., q_{N-1}\}, k \neq j$, listening to a Proposer j with strategy p_j, is given by

$$\Pi_R(p_j, q_{-j}) = \frac{p_j}{N-1} a_{p_j, q_{-j}} \quad (3)$$

Assuming that these games take place in groups where each individual acts once as a Proposer (in turns and following a round robin fashion), the overall payoff of an individual with strategy (p_i, q_i), playing in a group where opponent strategies are summarised in the strategy profile (p_{-i}, q_{-i}), is given by,

$$\Pi(p_i, q_i, p_{-i}, q_{-i}) = \Pi_P(p_i, q_{-i}) + \sum_{p_j \in p_{-i}} \Pi_R(p_j, q_{-j}) \quad (4)$$

The interesting values of M range between 1 and $N-1$. If $M < 1$ all proposals would be accepted and having $M > N-1$ would dictate unrestricted rejections. If $N = 2$ and $M = 1$, payoff function above reduces to

$$\Pi(p_1, q_1, p_2, q_2) = \begin{cases} 1 - p1 + p2, & \text{if } p1 \geq q2 \text{ and } p2 \geq q1. \\ 1 - p1, & \text{if } p1 \geq q2 \text{ and } p2 < q1. \\ p2, & \text{if } p1 < q2 \text{ and } p2 \geq q1. \\ 0, & \text{if } p1 < q2 \text{ and } p2 < q1. \end{cases} \quad (5)$$

recovering the traditional 2-person UG, described above [12, 18, 22].

MUG has interesting connections with typical N-person cooperation games, namely the ones with thresholds [24, 20, 25]. Indeed, defining altruistic cooperation as giving a benefit to the other incurring in a cost, we may say that a Proposer has a cost of p in order to provide a benefit of $p/(N-1)$ to the Responders. This way, fair proposals are cooperative gestures. Comparing with typical Public Good Games (PGG) with thresholds, in MUG *i)* we have a zero-sum game in which the multiplication factor is 1, promoting an unfavourable scenario for cooperation to thrive; *ii)* the threshold that dictates a successful proposal is endogenously imposed by each Responder; *iii)* individual offers, instead of group achievement, are the subject of suffrage and *iv)* the risk of failure, when a proposal does not comply with group threshold, is 1.

We further include a disagreement cost payed by the Proposer when her offer is rejected, that resembles an opportunity cost, the psychologic cost of having a proposal rejected or even the environmental cost of not reaching an agreement. When explicitly stated, this disagreement cost (d) affects Eq.(1) following

$$\Pi_P(p_i, q_{-i}) = (1 - p_i) a_{p_i, q_{-i}} - d(1 - a_{p_i, q_{-i}}) \quad (6)$$

2.1 Sub-game perfect equilibrium

To predict the outcome of the game previously introduced, we start by doing a typical equilibrium analysis. In this case, the predictions regarding Nash Equilibria (i.e., a strategy profile from which no player has interest in deviating alone) can be misleading, as those are well suited for non-sequential games. In sequential extensive form games, as MUG, the strategy profiles that are robust (i.e., that players looking forward to maximize utility will stick with) can be provided by the notion of the *sub-game perfect equilibrium* [19].

Let us first introduce some canonical notation. The game given in a sequential form has a set of stages in which a specific player (chosen by a *player function*) should act. A *history* stands as any possible sequence of actions, given the turns assigned by the player function. Roughly speaking, a *terminal history* is a sequence of actions that go from the beginning of the game until an end, after which there are no actions to follow. Each *terminal history* will prescribe different outcomes to the players involved, given a specific *payoff* structure that fully translates the preferences of the individuals.

A *sub-game* is (again, a game) composed by the set of all possible histories that may follow a given non-terminal history. Informally, a sub-game is the game yet to play, after a given sequence of actions already performed by the players. A strategy profile is a *sub-game perfect equilibrium* if it also the Nash equilibrium of every sub-game, i.e., a Nash equilibrium of the sub-games that follow any possible sequence of actions (non-terminal histories).

Let us turn to the specific example of MUG to clarify this idea. In this game, the *player function* dictates that the Proposer does the first move and, after that, the Responders should state acceptance or rejection. The game has two stages and any terminal history is composed by sets of two actions, one taken by a single individual (Proposer, that may suggest any division of the resource) and the second by the group (acceptance or rejection).

Picture the scenario in which groups consist in 5 players, where one is the Proposer, the other 4 are the Responders and M=4 (different M would lead to the same conclusions). Let us evaluate two possible strategy profiles: $s_1 = (0.8, 0.8, 0.8, 0.8, 0.8)$ and $s_2 = (\mu, 0, 0, 0, 0)$, where the first value is the offer by the Proposer and the remaining 4 are the acceptance thresholds by the Responders. Both strategy profiles are Nash Equilibria of the whole game. In the first case, the Proposer does not have interest in deviating from 0.8: if she lowers this value, the proposal will be rejected and thus she will earn 0; if she increases the offer, she will keep less to herself. The same happens with the Responders: if they increase the threshold, they will earn 0 instead of 0.2, and if they decrease it, nothing happens (non-strict equilibrium). The exact same reasoning can be made for s_2, assuming that $\mu/(N-1)$ is the smallest possible division of the resource.

Regarding sub-game perfection, the conclusions are different. Assume the *history* in which the Proposer has chosen to offer μ (let's call the sub-game after this history, in which only one move is needed to end the game, h). In this case, the payoff yielded by s_1 is $(0, 0, 0, 0, 0)$ (every Responder rejects a proposal of μ) and the payoff yielded by s_2 is $(1 - \mu, \mu/(N-1), \mu/(N-1), \mu/(N-1), \mu/(N-1))$. So it pays for the Responders to choose s_2 instead of s_1, which means that s_1 is not a Nash Equilibrium of the sub-game h. Indeed, while any strategy profile in the form $s = (p, p, p, p, p), \mu < p \leq 1$ is a Nash Equilibrium of MUG, only $s^* = (\mu, 0, 0, 0, 0)$ is the sub-game perfect equilibrium. As described in the introductory section, a similar conclusion, yet simpler and more intuitive, could be reached through backward induction.

This sub-game perfect equilibrium prescribes a payoff of $1 - \mu$ to the Proposer and μ/N to the Responder, therefore,

in terms of fairness, the scenario is dark. In real life, individuals do not play this way. Would artificial agents learn sub-game perfection or would they learn to behave fair as humans?

3. LEARNING MODEL

We use the Roth-Erev algorithm [23] to analyse the outcome of a population of learning agents playing MUG in groups of size N. In this algorithm, at each time-step t, each agent k is defined by a propensity vector $Q_k(t)$. This vector will be updated considering the payoff gathered in each play. This way, successfully employed actions will have high probability of being repeated in the future. We consider that games take place within a population of size $Z > N$ of adaptive agents. To calculate the payoff of each agent, we sample random groups without any kind of preferential arrangement (well-mixed assumption). We consider MUG with discretised strategies. We round the possible values of p (proposed offers) and q (individual threshold of acceptance) to the closest multiple of $1/D$, where D measures the granularity of the strategy space considered. We map each pair of decimal values p and q into an integer representation, thereafter $i_{p,q}$ is the integer representation of strategy (p, q) and p_i (or q_i) designates the p (q) value corresponding to the strategy with integer representation i.

The core of the learning algorithm takes place in the update of the propensity vector of each agent, $Q(t+1)$, after a play at time-step t. Denoting the set of possible actions by $A, a_i \in A : a_i = \{p_i, q_i\}$ and the population size by Z, the propensity matrix, $Q(t) \in \mathbb{R}_+^{Z\times|A|}$ is updated following the base rule

$$Q_{ki}(t+1) = \begin{cases} Q_{ki}(t) + \Pi(p_i, q_i, p_{-i}, q_{-i}) & \text{if } k \text{ played } i \\ Q_{ki}(t) & \text{otherwise} \end{cases} \tag{7}$$

The above update can be enriched with human learning features: *forgetting rate* ($\lambda, 0 \leq \lambda \leq 1$) and *local experimentation*, ($\epsilon, 0 \leq \epsilon \leq 1$) [23], leading to an update rule slightly improved,

$$Q_{ki}(t+1) = \begin{cases} Q_{ki}(t)\bar{\lambda} + \Pi(p_i, q_i, p_{-i}, q_{-i})(1-\epsilon) & k \text{ played } i \\ Q_{ki}(t)\bar{\lambda} + \Pi(p_i, q_i, p_{-i}, q_{-i})\frac{\epsilon}{4} & k \text{ pl. } i_p \pm 1 \\ Q_{ki}(t)\bar{\lambda} + \Pi(p_i, q_i, p_{-i}, q_{-i})\frac{\epsilon}{4} & k \text{ pl. } i_q \pm 1 \\ Q_{ki}(t)\bar{\lambda} & \text{otherwise} \end{cases} \tag{8}$$

where $\bar{\lambda} = 1 - \lambda$ and $i_p \pm 1 (i_q \pm 1)$ corresponds to the index of the p (q) values of the strategies adjacent to p_i (q_i), naturally depending on the discretisation chosen. The introduction of local experimentation errors is convenient as they prevent the probability of playing less used strategies (however close to the used ones) from going to 0. Moreover, those errors may introduce the spontaneous trial of novel strategies, a feature that is both human-like and showed to improve the performance of autonomous agents [26]. The forgetting rate is convenient to inhibit the entries of Q to grow without bound: when the propensities reach a certain value, the magnitude of the values forgotten, $Q_{ki}(t)\lambda$, approach those of the payoffs being added, $\Pi(p_i, q_i, p_{-i}, q_{-i})$. All together, the individual learning algorithm can be intuitively perceived: when individual k uses strategy i she will reinforce the use of that strategy provided the gains that she obtained; higher gains will increase more the probability of using that strategy in the future. The past use of the remaining strategies, and the obtained feedbacks, will be forgotten over time; The similar strategies to the one employed (which in the case of MUG are just the adjacent values of proposal and acceptance threshold) will also be reinforced, yet to a lower extent.

When an agent is called to pick an action, she will do so following the probability distribution dictated by her normalised propensity vector. The probability that individual k picks the strategy i at time t is given by

$$\rho_{ki}(t) = \frac{Q_{ki}(t)}{\sum_n Q_{ni}(t)} \tag{9}$$

The initial values of propensity, $Q(0)$, have a special role in the convergence to a given propensity vector and on the exploration *versus* exploitation dilemma. If the norm of propensity vectors in $Q(0)$ is high, the initial payoffs obtained will have a low impact on the probability distribution. Oppositely, if the norm of propensity vectors in $Q(0)$ is small, the initial payoffs will have a big impact on the probability of choosing the corresponding strategy again. Convergence will be faster if the values in $Q(0)$ are low, yet in this case agents will not initially explore a wide variety of strategies.

Additionally, we consider a modified probability distribution that takes the form of a Gibbs-Boltzmann probability distribution. This distribution will be useful to introduce negative payoffs, occurring when we include disagreement costs (see Section 2).

$$\rho_{ki}(t) = \frac{e^{Q_{ki}(t)/\tau}}{\sum_n e^{Q_{kn}(t)/\tau}} \tag{10}$$

Parameter τ corresponds to a temperature: low values will highlight the differences in propensity values in the corresponding probability distribution, while high values will introduce stochasticity by softening the effect of the propensities on the probability of choosing a given action.

Algorithm 1: Roth-Erev reinforcement learning algorithm in an adaptive population and considering synchronous update of propensities.

$Q(0) \leftarrow$*random initialisation*;
for $t \leftarrow 1$ **to** *T, total number of time-steps* **do**
  $tmp \leftarrow \{0, ..., 0\}$ `/* keeps the temporary payoffs of the current generation to allow for synchronous update of propensities */`;
  for $k \leftarrow 1$ **to** Z **do**
    1. *pick random group with individual* k;
    2. *collect strategies (Eq. 9,10)*;
    3. *calculate payoff of* k *(Eq. 4)*;
    4. *update* $tmp[k]$ *with payoff obtained*;
  update $Q(t)$ *given* $Q(t-1)$ *and tmp (Eq. 8)*;
  save $\bar{p}$ *(Eq. 11)*;
  save $\bar{q}$ *(Eq. 11)*;

As said, we consider a population of Z learning agents. Propensities will be synchronously updated after each time-step (t). In a time-step, every agent plays once in a randomly assembled group. A general view over the learning algorithm is provided in Algorithm 1. After each t, we keep track of the average values of p and q in the population, designating them by $\bar{p}$ and $\bar{q}$. Provided a propensity matrix, they are calculated as

$$\bar{p} = \frac{1}{Z} \sum_{1<k<Z} \sum_{1<i<|A|} \rho_{ki} p_i$$

$$\bar{q} = \frac{1}{Z} \sum_{1<k<Z} \sum_{1<i<|A|} \rho_{ki} q_i \quad (11)$$

The learning algorithm employed is rather popular [8, 23], providing a representative form of individual based learning. Other algorithms, such as Q-learning [31, 3], Learning Automata [17, 6] or Cross Learning [4, 1], can be similarly employed [2]. In the scope of this work, a simple *stateless* formulation Q-learning can be used [3], whereby the update of propensities follows the rule

$$Q_{ki}(t+1) = \begin{cases} Q_{ki}(t) + \alpha(\Pi - Q_{ki}(t)) & \text{if } k \text{ played } i \\ Q_{ki}(t) & \text{otherwise} \end{cases} \quad (12)$$

where α stands for the learning rate and Π is used as a simplification for $\Pi(p_i, q_i, p_{-i}, q_{-i})$. Learning Automata implies the direct update of the own action usage probabilities (instead of updating an intermediary propensity vector). Using this method, the probabilities of using each strategy a are updated, from t to $t-1$, according to

$$\rho_{ki}(t+1) = \begin{cases} \rho_{ki}(t) + \alpha\Pi(1 - \rho_{ki}(t)) & \text{if } k \text{ played } i \\ \rho_{ki}(t) - \alpha\Pi\rho_{ki}(t) & \text{otherwise} \end{cases} \quad (13)$$

A comparison between each of these algorithms, in the context of autonomous agents interacting through MUG, is currently under progress.

4. RESULTS AND DISCUSSION

Through the simulation of the multiagent system described in the previous section, we first show that different group decision thresholds have a considerable impact on the average values of offers (p) and acceptance thresholds (q) learned by the population. As the time-series in Figure 1 show, both for $M = 1$ and $M = 4$ (the extreme cases when the group size is 5), agents learn the strategies that allow them to maintain high acceptance rates and high average payoffs. Notwithstanding, the offered values when $M = 4$ are fairer than the ones learned when $M = 1$. An average p of 0.2 ($M = 1$) endows Proposers with an average payoff of 0.8, while each Responder keeps 0.05. Oppositely, an average value of p close to 0.6 provides the equalitarian outcome of endowing Proposers with 0.4 and Responders with 0.15. If one assumes that the role of Responder will be played $(N-1)$ times often, then Responders earn 0.2 for $M = 1$ and 0.6 for $M = 4$ and here indeed, the group decision criteria is enough to even provide an advantage for Responders. Recall that sub-game perfect equilibrium always predicts that Proposers would keep all the sum and Responders would earn 0.

To have a better intuition for the distribution of strategies within a population, we take a snapshot, for a specific run, of

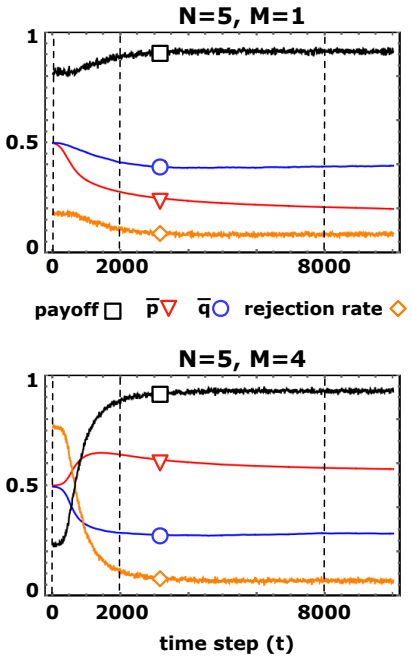

Figure 1: Time series reporting the evolution of average strategies ($\bar{p}$ and $\bar{q}$), average payoff population-wide and proposals rejection rate. Each plot corresponds to the average over various runs, each starting with a random propensity matrix where each entry is sampled from a uniform distribution from 0 to $Q(0)_{max}$. For group size $N = 5$ and for the extreme cases of threshold M ($M = 1$, $M = 4$), the rejection rate converges to a value near the minimum, thereby, the average payoff in the population approximates the maximum possible. Other parameters: population size $Z = 50$, granularity $D = 20$, forgetting rate $\lambda = 0.01$, local experimentation rate $\epsilon = 0.01$, total number of time-steps $T = 10000$, number of runs $R = 50$, disagreement cost $d = 0$, initial propensities maximum $Q(0)_{max} = 50$.

the population distribution over the space of possible p and q values for time-steps $t = 0$, $t = 2000$ and $t = 8000$. The corresponding results are pictured in Figure 2. Each square corresponds to a pair (p, q) and a darker square means that more agents have a propensity vector whose average strategy stands in that position. Figure 2 shows that, over time, agents learn to use a p value that grows with M. Concerning q, the learned values have a sizeable variance within the same population. This variance decreases with M. The reasoning for this result is straightforward: as M increases, a proposal is only accepted if more Responders accept it. In the limiting case of $M = N - 1$, all Responders have to accept an offer in order for it to be accepted by the group, thereby, the pressure for having low acceptance thresholds, q, is high. When M is low, a lot of q in the group of Responders turn to be irrelevant. If $M = 1$, a single Responder is enough to accept a proposals and thereafter, all the other q values in the group do not need to be considered. In this case, the pressure for q values to converge to confined domain is softened.

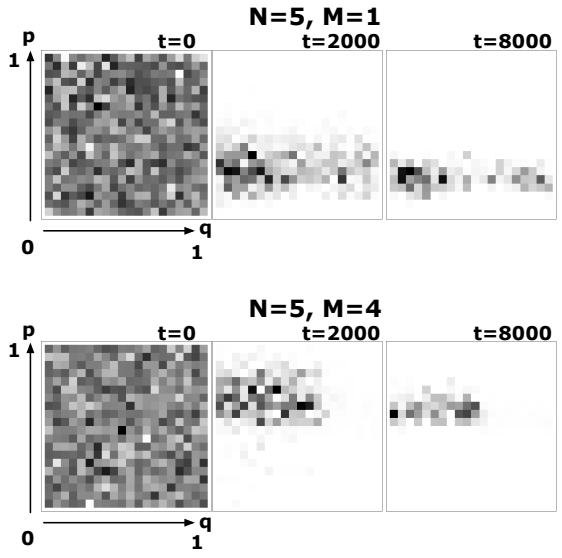

Figure 2: Snapshots of the population composition regarding the average values of p and q to be played given $Q(t)$. Each plot represents the space of all possible combination of p and q, assuming that $D = 20$ and thereby, p and q are rounded to the closest multiple of $1/D$. We represent the state of the population for three distinct time-steps ($t = 0$, $t = 2000$ and $t = 8000$) and given two values of threshold M, $M = 1$ and $M = 4$ (group size $N = 5$). The time location of these snapshots is represented in Figure 1 by means of vertical dashed lines. Each square within the 2D-plots represents a specific combination of (p, q). If the square is darker it means that more individuals of the population play, on average, with a strategy corresponding to that location. For accessing other fixed parameters, see the caption of Figure 1.

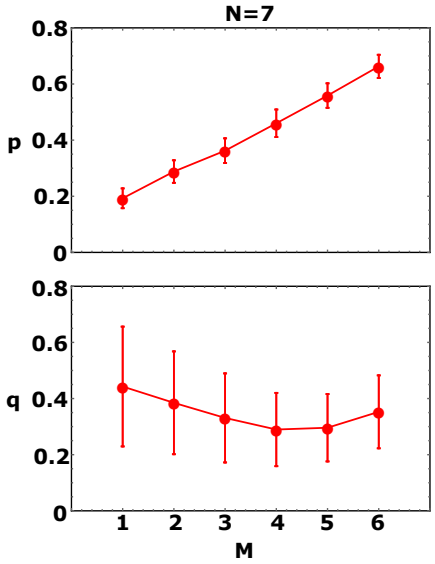

Figure 3: The average values of p and q for group size $N = 7$ with M assuming all possible non-trivial values $1 \leq M \leq N - 1$. Each point corresponds to a time and ensemble average: i) time average over the last half of the time-steps, i.e., we wait for a transient time for propensity values to stabilise and ii) we take the average of 50 runs, each one starting from a random $Q(0)$ propensity matrix. For other parameters, see caption of Figure 1.

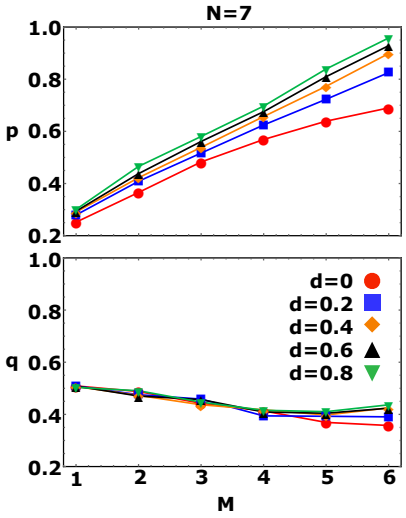

Figure 4: The effect of disagreement const, d, on the adopted values of p and q, for different M. Due to the possibility of having negative payoffs, this is the only scenario where the probabilities of selecting a given action are given by Eq. (10) instead of Eq. (9). We used $\tau(t) = \tau/t$ and $\tau = 10^4$. For other parameters, see caption of Figure 1.

The relation between M and within population strategy variance is further evidenced in Figure 3. Here we plot the average values of p and q, taken as the time average after a transient period of half of the total time-steps, T. The error bars represent the average (over time) of the standard deviation of the p and q values within the population. The standard deviation of q is clearly high and it decreases with M. Also here, the effect of stringent group acceptation criteria is evident, in what concerns the learning of being a fair Proposer.

The effect of M can even be leveraged if we include disagreement costs (d). As Figure 4 shows, increasing the cost that a Proposer incurs in when the quorum of Responders rejects a proposal has the effect of increasing the values proposed. Once again, if we followed the prediction stemming from sub-game perfection (Section 2.1) we would not take into account the possible effects of a disagreement cost. If we considered that all the proposals were to be accepted by the Responders, the Proposer would never fear the disagreement cost, and this parameter would be innocuous.

Finally, we highlight the effect of group size (N), on the average value of proposals made and proposals willing to be accepted. As Figure 5 depicts, larger groups induce individuals to rise their average acceptance threshold. It is reasonable to assume that, as the group of Responders grows

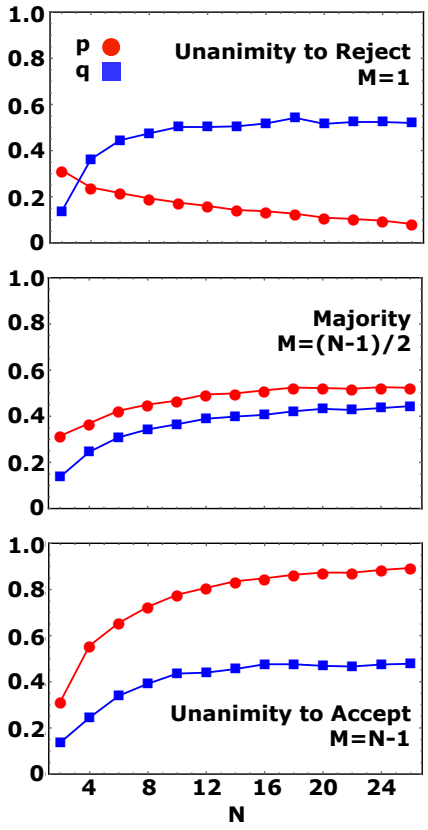

Figure 5: Average values of p and q for different combinations of group sizes, N, and group decision criteria, M. Other parameters: population size $Z = 50$, granularity $D = 20$, forgetting rate $\lambda = 0.01$, local experimentation rate $\epsilon = 0.01$, total number of time-steps $T = 10000$, number of runs $R = 50$, disagreement cost $d = 0$, initial propensities maximum $Q(0)_{max} = 50$.

and as they have to divide the offers between more individuals, the pressure to learn optimal low q values is alleviated. This way, the values of q should increase, on average, approaching the 0.5 barrier that would be predicted if they behaved erratically. Differently, the proposed values exhibit a dependence of the group size that is conditioned on M. For mild group acceptance criteria (low M), having a big group of Responders is synonym of having a proposal easily accepted. In these circumstances, Proposers tend to offer less without risking having their proposals rejected, keeping this way more for themselves and exploiting the Responders. Oppositely, when groups agree upon stricter acceptance (values of M that, as Figure 5 shows, can go from majority to unanimity), having a big group of Responders means that more persons need to be convinced of the advantages of a proposal. This way, Proposers have to adapt, increase the offered values and sacrifice their share in order to have their proposals accepted. We tested these results for values of local experimentation error (ϵ) and forgetting rate (λ) in the set $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. While high values of ϵ and λ lead to a slight decrease in the average values of p and increase in q, the conclusions regarding the effects of M, N and d remain the same. We additionally tested for $N = 7$, $M = 1, 3, 6$ and $Z = 20, 30, 50, 100, 200, 300, 500$ and verified that the conclusions regarding the effect of M remain valid for the considered population sizes.

5. CONCLUSION

We are all part of large multi-agent systems and our preferences are (also) the result of adaptation and response within those systems. The path taken during that adaptive process may disembogue in nontrivial behaviors. Being fair is an example. Why and how did we end up being fair are questions that may never be fully answered, however, trying to do so turns to be paramount if we want to understand societies and design fruitful institutions. The mathematical or computational apprehension of fairness turns to be extremely relevant in the contemporary digital societies. More than being part of human multi-agent systems, we are today interacting with artificial agents. Take the example of automatic negotiation [14, 15]. What would be the requirements of artificial agents designed to negotiate with a human in an environment that is surely dynamic? Should they behave assuming human rationality and predicting sub-game perfect equilibrium (see Section 2.1)? Should they learn with the dynamics of the environment and opponents?

We employ a reinforcement learning algorithm to shed light on the role of decision rules, group size and disagreement costs. We model an adaptive population in which learning agents shape both their propensities and ergo, opponents' playing environment. We show that increasing the group acceptation threshold has the effect of increasing the offered values and decreases the acceptance thresholds. The imposition of disagreement costs, to be paid by the Proposers in case of having a proposal rejected, even helps to leverage group fairness. Moreover, the effect of group size depends on the group decision rule: big groups combined with soft group criteria are a fertile ground for selfish Proposers to thrive. Oppositely, big groups that require unanimity to accept a proposal, by being strict in accepting low proposals, induce Proposers to offer more.

The individual learning model that we implement is close to a trial and error mechanism that individuals may use to successively adapt to the environment, given the feedback provided by their own actions. A different approach implements a system of social learning [25], in which individuals learn by observing the strategies of others and accordingly imitate the strategies perceived as best. These two learning paradigms (individual and social) can lead to very different outcomes, concerning the learned strategies and the long-term behaviour of the agents [32]. Interestingly, our results (besides providing new intuitions regarding the role of disagreement costs and group size in MUG) are in line with some of the results obtained in the context of evolutionary game theory and social learning [25].

Acknowledgments

This research was supported by Fundação para a Ciência e Tecnologia (FCT) through grants SFRH/BD/94736/2013, PTDC/EEI-SII/5081/2014, PTDC/MAT/STA/3358/2014 and by multi-annual funding of CBMA and INESC-ID (under the projects UID/BIA/04050/2013 and UID/CEC/50021/2013 provided by FCT).

6. REFERENCES

[1] T. Börgers and R. Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14, 1997.

[2] D. Catteeuw, B. Manderick, S. Devlin, D. Hennes, and E. Howly. The limits of reinforcement learning in lewis signaling games. In *Proceedings of the 13th Adaptive and Learning Agents Workshop*, pages 22–30, 2013.

[3] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI/IAAI*, pages 746–752, 1998.

[4] J. G. Cross. A stochastic learning model of economic behavior. *The Quarterly Journal of Economics*, pages 239–266, 1973.

[5] S. De Jong, K. Tuyls, and K. Verbeeck. Artificial agents learning human fairness. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 863–870. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[6] S. de Jong, S. Uyttendaele, and K. Tuyls. Learning to reach agreement in a continuous ultimatum game. *Journal of Artificial Intelligence Research*, pages 551–574, 2008.

[7] R. Duch, W. Przepiorka, and R. Stevenson. Responsibility attribution for collective decision makers. *American Journal of Political Science*, 59(2):372–389, 2015.

[8] I. Erev and A. E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, pages 848–881, 1998.

[9] U. Fischbacher, C. M. Fong, and E. Fehr. Fairness, errors and the power of competition. *Journal of Economic Behavior & Organization*, 72(1):527–545, 2009.

[10] R. Forsythe, J. L. Horowitz, N. E. Savin, and M. Sefton. Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347–369, 1994.

[11] B. Grosskopf. Reinforcement and directional learning in the ultimatum game with responder competition. *Experimental Economics*, 6(2):141–158, 2003.

[12] W. Güth, R. Schmittberger, and B. Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388, 1982.

[13] E. Hoffman, K. McCabe, and V. L. Smith. Social distance and other-regarding behavior in dictator games. *The American Economic Review*, pages 653–660, 1996.

[14] N. R. Jennings, P. Faratin, A. R. Lomuscio, S. Parsons, M. J. Wooldridge, and C. Sierra. Automated negotiation: prospects, methods and challenges. *Group Decision and Negotiation*, 10(2):199–215, 2001.

[15] R. Lin and S. Kraus. Can automated agents proficiently negotiate with humans? *Communications of the ACM*, 53(1):78–88, 2010.

[16] D. M. Messick, D. A. Moore, and M. H. Bazerman. Ultimatum bargaining with a group: Underestimating the importance of the decision rule. *Organizational Behavior and Human Decision Processes*, 69(2):87–101, 1997.

[17] K. S. Narendra and M. A. Thathachar. *Learning automata: an introduction*. Courier Corporation, 2012.

[18] M. A. Nowak, K. M. Page, and K. Sigmund. Fairness versus reason in the ultimatum game. *Science*, 289(5485):1773–1775, 2000.

[19] M. J. Osborne. *An Introduction to Game Theory*. Oxford University Press New York, 2004.

[20] J. M. Pacheco, F. C. Santos, M. O. Souza, and B. Skyrms. Evolutionary dynamics of collective action in n-person stag hunt dilemmas. *Proceedings of the Royal Society B: Biological Sciences*, 276(1655):315–321, 2009.

[21] K. M. Page and M. A. Nowak. Empathy leads to fairness. *Bulletin of Mathematical Biology*, 64(6):1101–1116, 2002.

[22] K. M. Page, M. A. Nowak, and K. Sigmund. The spatial ultimatum game. *Proceedings of the Royal Society of London B: Biological Sciences*, 267(1458):2177–2182, 2000.

[23] A. E. Roth and I. Erev. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8(1):164–212, 1995.

[24] F. C. Santos and J. M. Pacheco. Risk of collective failure provides an escape from the tragedy of the commons. *Proceedings of the National Academy of Sciences*, 108(26):10421–10425, 2011.

[25] F. P. Santos, F. C. Santos, A. Paiva, and J. M. Pacheco. Evolutionary dynamics of group fairness. *Journal of Theoretical Biology*, 378:96–102, 2015.

[26] P. Sequeira, F. S. Melo, and A. Paiva. Emergence of emotional appraisal signals in reinforcement learning agents. *Autonomous Agents and Multi-Agent Systems*, 2014.

[27] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT press Cambridge, 1998.

[28] A. Szolnoki, M. Perc, and G. Szabó. Defense mechanisms of empathetic players in the spatial ultimatum game. *Physical Review Letters*, 109(7):078701, 2012.

[29] L. Tesfatsion. Agent-based computational economics: Growing economies from the bottom up. *Artificial Life*, 8(1):55–82, 2002.

[30] R. H. Thaler. Anomalies: The ultimatum game. *The Journal of Economic Perspectives*, pages 195–206, 1988.

[31] K. Tuyls, K. Verbeeck, and T. Lenaerts. A selection-mutation model for q-learning in multi-agent systems. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 693–700. ACM, 2003.

[32] S. Van Segbroeck, S. De Jong, A. Nowé, F. C. Santos, and T. Lenaerts. Learning to coordinate in complex networks. *Adaptive Behavior*, 18(5):416–427, 2010.

[33] S. Van Segbroeck, J. M. Pacheco, T. Lenaerts, and F. C. Santos. Emergence of fairness in repeated group interactions. *Physical Review Letters*, 108(15):158104, 2012.