



Computer Modeling of Cooperation and Altruistic Punishment

GIANGIACOMO BRAVO

Dipartimento di Studi Sociali, Università di Brescia

gbravo@eco.unibs.it

LUCIA TAMBURINO

Interdisciplinary Research Institute on Sustainability

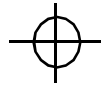
lucia.tamburino@unito.it

Abstract

Understanding cooperation is a long-standing problem in both biological and social sciences. Empirical evidence in repeated public-good (PG) experiments shows that intermediate levels of cooperation at the beginning of the game are followed by a decline over time ending close to the non-cooperative equilibrium. On the other hand, recent experiments have shown that introducing the possibility of punishing the defectors in PG experiments leads to high cooperation levels notwithstanding the fact that the punishment itself constitutes a second-order PG dilemma. This, and similar, results can be modeled using different types of actors. In this paper we use computer simulations to show that, using a plausible distribution of agent types whose behavior is related respectively to rational egoism (self interest), simple reciprocity and strong reciprocity schemes, it is possible to reproduce the results of PG experiments with a fair degree of accuracy at both the macro (aggregate) and the micro (individual) levels. Rational choice, simple reciprocity and strong reciprocity schemes are especially important because of their role in evolutionary theories. This outcome confirms the role of both simple and strong reciprocity in the understanding of the evolution of human behavior. Moreover, it advocates in favor of strong reciprocity as a fundamental scheme to understand a number of human behaviors, including altruistic punishment.

Keywords: cooperation, public good experiment, strong reciprocity, simulation





Understanding cooperation is a long-standing problem in both biological and social sciences (e.g. Hammerstein, 2003). Empirical evidence in repeated public good (PG) experiments shows that intermediate levels of cooperation at the beginning of each session are followed by a decline over time of the contributions to the public good, ending close to the non-cooperative equilibrium (Ledyard, 1995). Ernst Fehr and Simon Gächter (2000 and 2002), following Elinor Ostrom and colleagues' seminal paper (1992), showed that introducing the possibility of punishing the defectors in PG experiments leads to high cooperation levels notwithstanding the fact that the punishment itself constitutes a second-order PG dilemma (Yamagishi, 1986). This (and similar) result can be modeled using heterogeneous types of actors (Ahn, *et al.*, 2003b; Burlando & Guala, 2005; Kurzban & Houser, 2005). In this paper we show that - using a plausible distribution of agent types whose behavior is related respectively to rational egoism, simple reciprocity and strong reciprocity schemes - it is possible to simulate the results of PG experiments with a fair degree of accuracy at both the macro (aggregate) and the micro (individual) level. Rational egoism, simple reciprocity and strong reciprocity schemes are especially important because of their role in evolutionary theories (Bowles & Gintis, 2004; Fehr & Fischbacher, 2003; Gintis, *et al.*, 2003; Hammerstein, 2003). This outcome confirms the role of both simple and strong reciprocity in the understanding of the evolution of human behavior.

In the next section, we will briefly summarize Fehr and Gächter's experimental design and their main results. The following one presents a possible interpretation of the empirical results by using multiple actor types. The third section describes our model while the fourth shows the simulation results. The final one is devoted to the discussion.

1. The Original Experiments

Although any empirical interaction situation may, in principle, be reproduced by agent-based simulation, we chose to concentrate on Fehr and Gächter's (2000; 2002) experiments due to their representativity of general PG studies, the large number of subjects involved¹ and the important insights they offer by introducing of a condition where spontaneous and decentralized punishment is possible. In Fehr and Gächter's (2000; 2002) experiments, subjects play a repeated PG game for, respectively, twenty and twelve periods in groups of $n = 4$ individuals. In every period, each subject has an endowment of 20 monetary units (MUs) and chooses how many of them to contribute to a common project (the public good). The period payoff for each subject i is given by

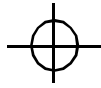
$$(1) \quad \pi_i = 20 - c_i + 0.4 \sum_{j=1}^n c_j$$

where c_i is the contribution of subject i to the project, 0.4 represents the marginal *per capita* return from the contribution and c_j are the contributions of all the members of i 's group (including i). During half of the experiment no punishment is allowed, while in the other half participants can impose a fee to the other members of their group at a cost for themselves.² The experiment

¹ A total of 352 subjects participated in the two Fehr and Gächter's experiments, 112 in the (2000) one and 240 in the (2002) one.

² Half of the subjects played the punishment condition first while the other half started playing the no-punishment condition in order to control for spillover effects between the two conditions.





has a two-order Prisoner's dilemma (PD) structure of interaction and the dominant strategy for all participants is neither to contribute to project nor to punish free-riders.³

As in most repeated PG experiment, Fehr and Gächter's results approached (without actually reaching it) the zero-contribution equilibrium prediction in the final period of the no-punishment condition: the mean contribution rapidly decreased from about 50% of the participants' endowment in the first period to less than 20% in the last one. When the punishment option was available, however, many subjects seized the possibility of sanctioning non cooperators notwithstanding the fact that punishment constitutes itself a PG dilemma (Yamagishi, 1986). This led to a very different outcome: the steady increase of average cooperation levels up to 70-90% of the optimum.

At the micro (individual) level the difference between the punishment and the no-punishment conditions is also striking. While 3/4 of the subjects played the dominant zero-contribution strategy in the final period of the punishment condition, in the no-punishment condition individual contributions in the final period were more scattered, but most above half of the endowment, with highest contribution frequencies at the 12, 15 and 20 MU level (Fehr & Gächter, 2000).

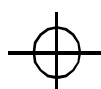
2. A Possible Interpretation of Experimental Results

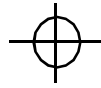
Simple rational choice theory has a hard time in explaining contribution in PG games while a theory that encompasses multiple type of players looks more suitable for the task (Burlando & Guala, 2005; Fehr & Gächter, 2002; Gintis, *et al.*, 2003; Ostrom, 2000). Both the initial high contribution levels and their decline during the no-punishment periods may be explained assuming that a proportion p of subjects playing a repeated PG game are rational egoists players, i.e. self-interested rational subjects (Ostrom, 2000), while a proportion q are simple reciprocity players, where $p + q = 1$. Actors modeled as rational egoist are payoff-maximizers: in PD-type situations (including PG games), they follow a dominant strategy of defection. Their subjective payoff preference order is alike the objective one: $\pi(I \text{ defect}, \text{others cooperate}) > \pi(I \text{ cooperate}, \text{others cooperate}) > \pi(I \text{ defect}, \text{others defect}) > \pi(I \text{ cooperate}, \text{others defect})$.

On the other hand, simple reciprocators are actors willing to cooperate in the first period of interaction. In the following ones, they respond with cooperation to the cooperation and with defection to the defection of other actors who interact in the same situation. Their subjective payoff preference order is $\pi(I \text{ cooperate}, \text{others cooperate}) > \pi(I \text{ defect}, \text{others cooperate}) > \pi(I \text{ defect}, \text{others defect}) > \pi(I \text{ cooperate}, \text{others defect})$. In PG situations without punishment, simple reciprocators start by contributing a substantial proportion of their endowment in the first period, but retaliate against the free-rider behavior of rational egoists by reducing their contribution in the following ones. The reduction of contribution carried out by the simple reciprocators in reaction to free-riding is consistent with the empirically observed gradual decrease of contribution levels over time (Gintis, *et al.* 2003; Ostrom, 2000).

Any model including only rational egoist and simple reciprocators shows some difficulties in fully interpret the empirical results presented above. More specifically, it appears unable to

³ At least in the sessions that used a stranger matching protocol, i.e. where the group composition change after every period.





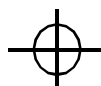
explain (i) why cooperation approaches, but never reaches zero in no-punishment periods and (ii) why so many subjects punish the free-riders in the periods where it is allowed.

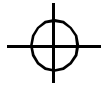
- (i) Since rational egoist never cooperate and simple reciprocators gradually reduce their contributions, cooperation level should reach zero in due time. Nevertheless, there is little evidence of this extreme decline. Most of the times, average contribution levels in the final round of repeated PG-games lie between 10% and 20% of the endowment and the proportion of subjects who contribute nothing does not exceed 80-90% (Fehr & Gächter, 2000; Ledyard, 1995; Ostrom, 2003). Moreover, when subjects know in advance the number of repetitions of the game, the rate of decay of cooperation is inversely related to the number of played periods and even after 60 repetitions cooperation levels remain above 10% (Isaac, *et al.*, 1994).
- (ii) The understanding of the reasons behind the widespread punishments is not straightforward in the light of a model that encompasses only rational egoists and simple reciprocators. Fehr and Gächter's (2002) experimental data shows that 84.3% of the 240 subjects punished at least once during the six periods of the punishment condition, 34.3% of the subjects punished more than five times and 9.3% punished more than ten times. Rational egoists, following their dominant strategy of defection, never punish while simple reciprocators are conditionally cooperative on punishment,⁴ just as they are on contribution. By considering a population of players composed by only those two types we should, therefore, observe a decrease over time of the second-order PG "credible threat of sanctions against the free-riders", just as the first-order PG "group contribution" declines. Consequently, the aggregate result should be different from the empirically observed constant rise of cooperation levels during the punishment condition.

To summarize: a model that includes only rational egoists and simple reciprocators helps to understand the dynamics of PG games, but is insufficient to reach a satisfying explanation of the empirical findings. On the other hand, the inclusion in the model of a limited number of strong reciprocity actors (Fehr & Fischbacher, 2003; Gintis, 2000; Gintis, *et al.*, 2003) can significantly improve its explanatory capacity.

Strong reciprocators are actors willing both to cooperate and to pay a cost in order to punish the non-cooperators, even when it is unlikely that their cooperation will be reciprocated and that other subjects will also punish the free-riders. While strong reciprocators always choose to cooperate in PD-type situations, their motives for cooperation may however be mixed. If concerned with group welfare considerations (Fehr & Gächter, 2002; Gintis, *et al.*, 2003), they cooperate because their subjective payoff preference order is akin to $\pi(I \text{ cooperate, others cooperate}) > \pi(I \text{ cooperate, others defect}) > \pi(I \text{ defect, others cooperate}) > \pi(I \text{ defect, others defect})$. However, a similar behavior may also result from a number of different motivations, among whose costly signaling may play a peculiar role (Fehr & Gächter, 2002; Gintis, *et al.*, 2003). According to costly signaling theory, individuals may act altruistically in order to signal the possession of favorable, yet unobservable, traits to potential mates, allies or enemies. The cost

⁴ At least in the sessions that used a stranger matching protocol, i.e. where the group composition change after every period.





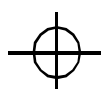
of the signal guarantees the receiver that the signal is honest (i.e. the signaler actually possesses the hidden traits), since otherwise the sender could probably not afford it (Gintis, *et al.*, 2001; Wright, 1999; Zahavi & Zahavi, 1977). A subject using costly signaling may therefore play cooperatively also in a context of universal defection in order to signal his/her hidden qualities. His/her subjective payoff preference order is $\pi(I \text{ cooperate, others defect}) > \pi(I \text{ cooperate, others cooperate}) > \pi(I \text{ defect, others defect}) > \pi(I \text{ defect, others cooperate})$. Either because of pro-social attitudes or of costly signaling opportunities, strong reciprocators are always willing to cooperate and to bear a personal cost in order to punish free-riders. If r is the proportion of strong reciprocator subjects among the participants of a repeated PG game, we have $p + q + r = 1$.

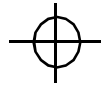
The introduction of strong reciprocators helps to overcome the two problems depicted above. (i) In the no-punishment condition, even when all simple reciprocators free-ride, strong reciprocators continue to contribute to the PG, accounting for the 10-20% of cooperation which lasts until the final period. (ii) In the punishment condition, their will to punish free-riders in spite of what others do is fundamental to assure the provision of the second-order PG “credible threat of sanctions against the free-riders” needed to achieve high contribution levels.

It is worth noting that the evolutionary significance of the behavioral schemes depicted above is well grounded in the light of some recent findings (Fehr & Fischbacher, 2003; Gintis, *et al.*, 2003; Hammerstein, 2003; Ostrom, 2000; Ostrom & Walker, 2003). Empirical evidence, both in the lab and in the field, as well as day-to-day experience show that Human cooperation extends far beyond what can be explained by kin selection theory (Hamilton, 1964) and even by reciprocal altruism theory (Trivers, 1971). In addition of kin selection and reciprocal altruism, evolutionary models that account for the evolution of strong altruist behaviors have recently been developed. They include group selection (Gintis, 2000), cultural evolution and gene-culture co-evolution (Richerson, *et al.*, 2003) and costly signaling through pro-social behaviors (Gintis, *et al.*, 2001) as mechanisms leading to the evolution of strong reciprocators. While it is impossible to illustrate here the details of those models, it is important to note that, taken together, they appear consistent with the experimental findings of PG games and may represent an important step in the direction of a better understanding of the evolution of human behavior.

In order to use rational egoism, simple reciprocity and strong reciprocity as basement stones for the building of a computer model, we need to have an idea of the distribution of the different actor types among the participants of repeated PG experiments. A number of studies tried to identify actor types using different methodologies. Reminding that p denotes the proportion of rational egoist, q the proportion of simple reciprocators and r the proportion of strong reciprocators, where $p + q + r = 1$, we will briefly present their results in order to fix empirically plausible intervals for the variation of p , q , and r . Those intervals will subsequently serve as rough boundaries for the variability of simulation parameters.

- John Ledyard (1995), in his review of PG experiments, suggest $p = 0.50$ and $q = 0.40$ plus a proportion of 0.10 subjects who “behave in an inexplicable (irrational?) manner”. However, he acknowledges that his estimation has no firm basis but is the result of “casual observation”.
- Urs Fischbacher and colleagues (2001) in a one-shot PG game asked the participants to fill a “contribution table”, indicating the contribution of each subject for any possible contribution of the other group members. This led to an estimation of $p = 0.30$ and $q = 0.50$,





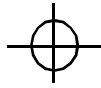
- plus a 0.14 of the subjects that presented a “hump-shaped” contribution schedule and a 0.06 with other patterns.
- Toh-Kyeong Ahn and colleagues (2003a), using questionnaire data, found that about 40% of subjects facing a one-shot PG game ranked the outcome (D,C) below the outcome (C,C) and the outcome (CD) below the outcome (D,D). They refer to those type of players with the expression “assurance-type” players. Using the p - q - r notation, they found $p = 0.58$, $q = 0.42$, and $r = 0.00$.
 - In a different paper, Toh-Kyeong Ahn and colleagues (2003b) analyze both the results of a class survey at Indiana University and the questionnaires answers given during Hayashi *et al.* (1999) U.S. experiment. The first study leads to $p = 0.42$ and $q = 0.29$, with a further 0.29 of subjects presenting other patterns. It is worth noting that, among those, $r = 0.12$ ranked the outcome (C,D) above the outcome (D,C), a result which is compatible with a strong reciprocity type of player. In the Hayashi *et al.* study, evidence suggests $p = 0.20$ and $q = 0.35$. A further 0.02 of subjects are labeled “strong altruists”, while 0.46 are labeled “other” or “not explained”, but, among those, 0.06 ranked the outcome (C,D) above the outcome (D,C), which leads to $r = 0.02 + 0.06 = 0.08$.
 - In a recent paper, Robert Kurzban and Daniel Houser (2005) propose $p = 0.20$, $q = 0.63$ and $r = 0.13$ (labeled “cooperators”), while 0.04 are “not classifiable”.
 - Using a set of four different methods, Roberto Burlando and Francesco Guala (2005) classified the subject participating to a repeated PG games in “free riders” ($p = 0.32$), “reciprocators” ($q = 0.35$) and “cooperators” ($r = 0.18$), plus a residual proportion of subjects (0.15) labeled as “noisy”.
 - In an (at present) unpublished work based on James Andreoni (1995) data, Riccardo Boero (pers. com.) found $p = 0.33$, $q = 0.35$ and $r = 0.25$ (labeled as “altruists”), while 0.07 were unclassifiable.

Notwithstanding the facts that the methodologies used for gathering the above data differ and that the possible interpretation of the data themselves are also manifold, it is possible to use those estimations to define some roughly credible intervals for the variation of the parameters regarding the proportion of actor types in the simulation runs. Bringing together the above figures, we have $p \in [0.20, 0.60]$, $q \in [0.30, 0.60]$ and $r \in [0.00, 0.25]$, with the further constraint $p + q + r = 1$. The next sections will therefore explore those intervals.

Summarizing, our argument is that rational egoism, simple reciprocity and strong reciprocity are three schemes, all consistent with evolutionary models, that social sciences can positively use to interpret situations of repeated interaction having a PG dilemma structure, even if this is not the only possible explanation of the empirical results.⁵ Our aim is to develop a model where those schemes are employed to implement the decision-making routines of computer agents. The main goal is to show that it is possible to simulate with a fair degree of approximation the experimental results of PG games akin to the Fehr and Gächter (2000) and (2002) ones by using the same distribution of agent types for both the no-punishment and the punishment conditions.

⁵ As recently suggested by James Fowler and colleagues (2005), egalitarian motives may play an important role in explaining the empirical results of PG games, including the differences in contributions between the punishment and the no-punishment conditions. Ernst Fehr and Klaus Schmidt (1999) advanced a similar argument in presenting their Inequity aversion model.





In addition, we will use the model in order to explore the distribution limits of agent types consistent with experimental evidence.

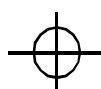
3. The Model

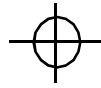
In designing the model, our aim was to reproduce as close as possible the structure of interaction of Fehr and Gächter's experiments. We used 240 agents⁶ that interacted in group of 4 for 20 periods: periods 1-10 are without punishment, while in periods 11-20 punishment is allowed. Groups are randomly re-matched at the beginning of each period. Agents are of three different types: type 1 models rational egoist actors; type 2 models simple reciprocator actors; type 3 models strong reciprocator actors. The distribution across types represents the main parameter of the model.

Each period starts with a routine aimed at randomly matching agents in groups. Afterwards, agents "decide" their own contribution to the PG. This is computed by using a function that returns a pseudo-random value from a normal distribution. The mean of the distribution depends on the agent's type. The distribution mean of Type 1 agent, its *center*, is equal to 0 (i.e. they do not cooperate) in the first period and does not change during the no-punishment condition. Type 2 agents start with a center equal to 20 (they fully cooperate) in the first period and, subsequently, they adapt their center to the mean contribution of their own group in the previous period. This simulates the fact that simple reciprocators start by playing cooperatively and respond by cooperating to the cooperation and by defecting to the defection of other subjects. Type 3 agent always cooperate and their centers are equal to 20. The standard deviation for all agent types is equal to 2, accounting for limited trembling hand effects. Since in the original experiment investments in the PG were limited between 0 and 20 MUs, all agent contribution greater than 20 are corrected to 20 whereas all contributions below 0 are corrected to 0. Afterwards, results for all agents are computed using the formula (1).

In period 11, the first period of the punishment condition, all agent centers are reset to their initial values. This simulates the fact that no significant behavioral differences has been empirically found between the groups playing first the no-punishment condition and the ones playing first the punishment condition. From period 11 onwards, a punishment stage is added to the previous ones: type 2 and type 3 agents have to decide whether to punish or not and the amount of the fee (type 1 agents never punish). If their contribution is larger than their group mean, they punish any other agent who has contributed less than the group mean. The amounts of the fees are computed by the formula: $f \times (\text{mean group contribution} - \text{punished agent contribution})$, where the parameter $f = 1.86$ represents the average fee per unitary deviation from the mean group contribution found by Fehr and Gächter in the (2002) experiment. For every MU used for punishing other agents, the punisher bears the cost of 1/3 MU. A fundamental effect of the punishment is that it modifies the punished agent center. In order to avoid further punishments, the punished agent adapt its center to the average of its group. This holds for all agent types.

⁶ 240 is the total number of subjects that participated in Fehr and Gächter 's (2002) experiment.





4. Results

By using a distribution of agent types corresponding to the “lower bound” of cooperative agents (i.e., according to the shares estimated above, $p = 0.60$, $q = 0.40$ and $r = 0.00$) the model weakly replicates Fehr and Gächter’s results (Fig. 1). At the macro (aggregate) level, the mean contribution starts at a level below 40% of the endowment and rapidly falls close to zero during the no-punishment condition while during the punishment one it slowly grows up reaching approximately 50% of the endowment. In both cases, cooperation levels appear to be lower than in the original experiments. At the micro (individual) level, close to 80% of agents do not contribute anything in the final period of the no-punishment condition while in the final period of the punishment condition contributions are almost normally distributed around 10 MUs, with an extremely low number of agents contributing either zero or 20.

On the other hand, the “upper bound” agent distribution ($p = 0.20$, $q = 0.55$ and $r = 0.25$) leads to over-cooperative outcomes (Fig. 2). The mean contribution start at a level over 75% of the optimum and slowly decrease during the no-punishment condition while, during the punishment condition, it ends above 90% of the optimum. The micro-level results, with 20 MUs as the modal contribution in the final period of both the no-punishment and the punishment conditions, are unlikely as well.

The outcomes above suggest that intermediate (although closer to the lower bound than to the upper one) distributions of agent types will lead to better replications of the empirical results. A first step may be to simply decrease the ratio p/q leaving $r = 0.00$ (i.e. to raise the number of simple reciprocators without introducing any strong reciprocator). This conducts to outcomes close to the empirical ones in the no-punishment condition while the increase of cooperation appears weaker than in the original experiments during the punishment condition. Also at the micro level, the distribution of contribution choices done by artificial agents in the final periods of the two conditions is much closer to the real subjects’ one for the no-punishment condition than for the punishment condition. The introduction of, at least, a limited number of type 3 agents appears therefore mandatory in order to appropriately simulate the dynamics of the punishment condition. This happens already with r as little as 0.05. However, the best replications of the experimental results are reached by using $p \cong 0.50$, $q \cong 0.40$ and $r \cong 0.10$ (Fig. 3). In this case, the trends for both the no-punishment and the punishment conditions are well defined and the magnitude of contribution changes is closest to the empirical data. Also at the micro level, the simulation outcomes approach the empirical results: in the final period of the no-punishment condition, a vast majority of agents contributes zero or close to zero; in the final period of the punishment condition, contributions are more scattered even if most agents contribute more than half of their endowments.

5. Discussion and Conclusions

Notwithstanding the fact that our model reproduces Fehr and Gächter’s findings with a fair degree of accuracy, a closer comparison between the simulation outcomes and the experimental results shows, at least, two important differences. First, the mean contribution of real subjects



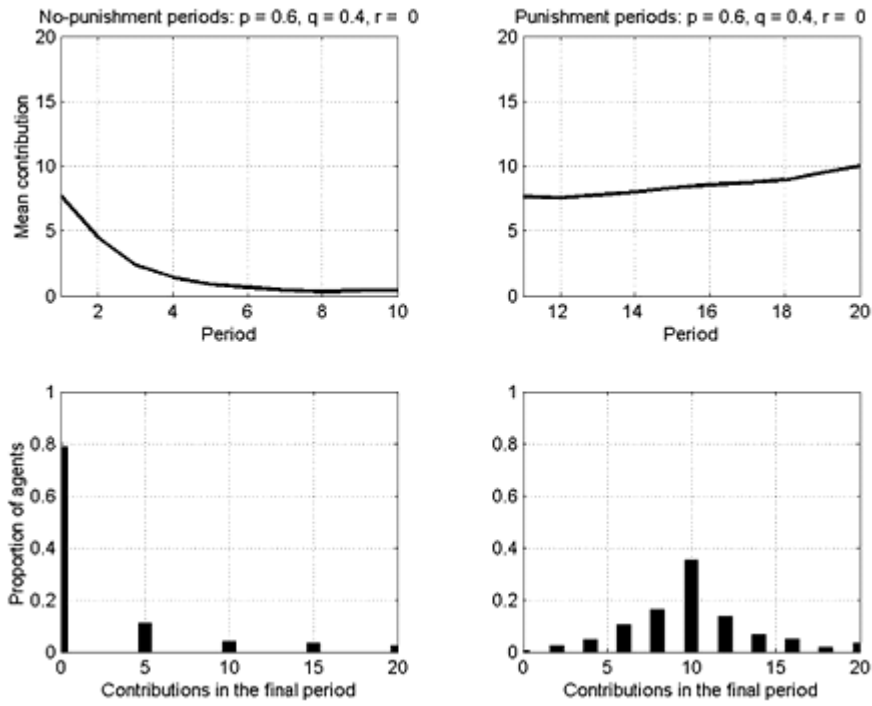
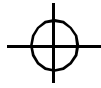


Figure 1 Contributions in a typical run with $p = 0.60$, $q = 0.40$ and $r = 0.00$.

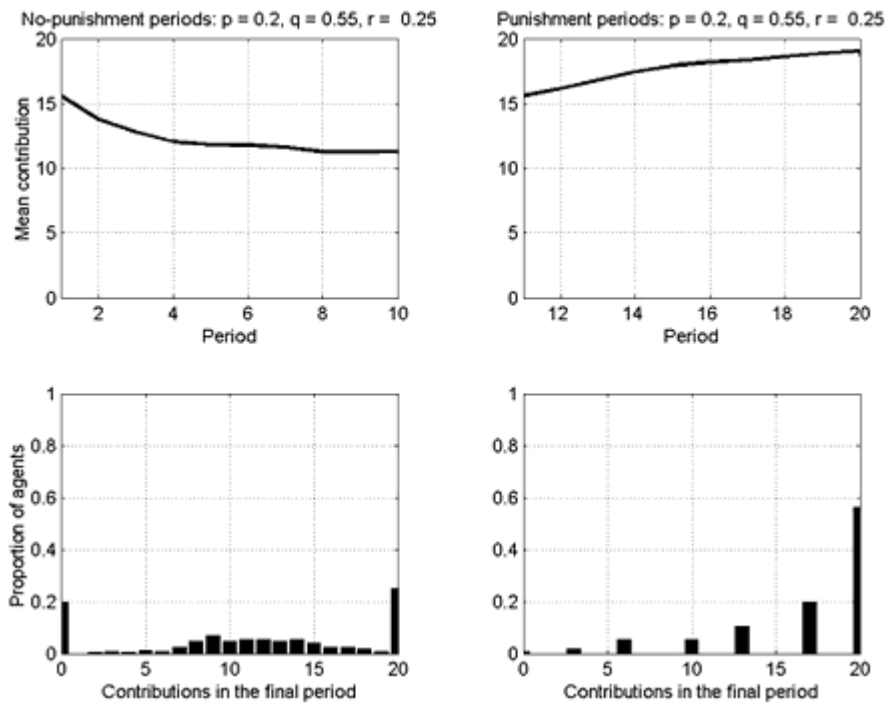
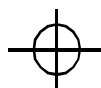


Figure 2 Contributions in a typical run with $p = 0.20$, $q = 0.55$ and $r = 0.25$.



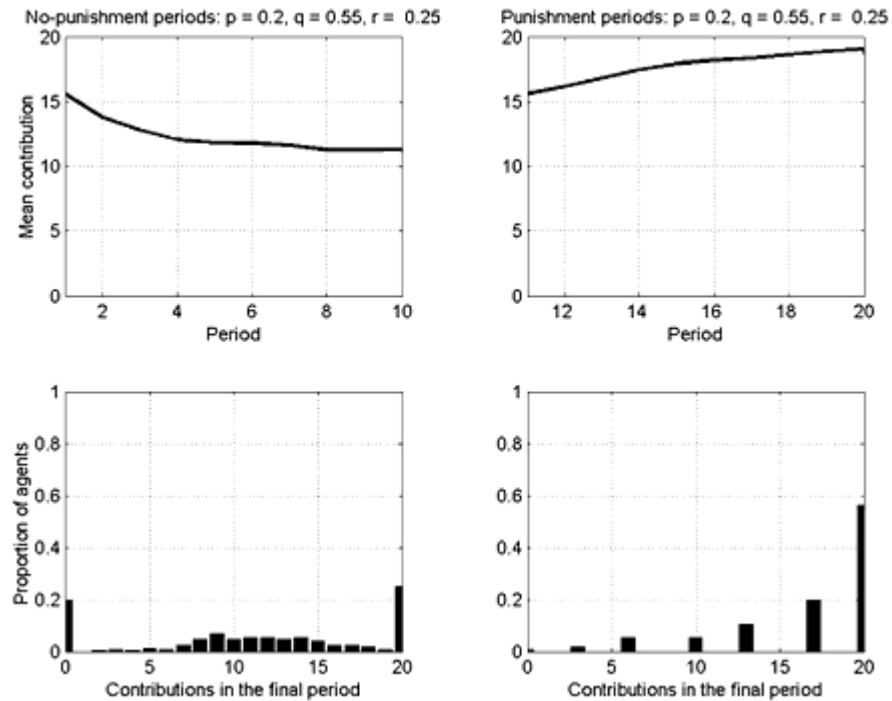
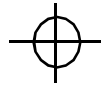


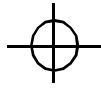
Figure 3 Contributions in a typical run with $p = 0.50$, $q = 0.40$ and $r = 0.10$.

in the starting period of the punishment condition is higher than the one in the starting period of the no-punishment one. This does not happen with artificial agents because, unlike human beings, our agents are not able to anticipate the fact that they would be punished if their contributions were below average: they need to be punished first in order to adapt their behavior. This is a clear limit of the cognitive capacities of the agents and a major difference with the behavior of humans.

Second, unlike real subjects, a limited number of agents still contribute their full endowment in the final period of the no-punishment condition. This holds for any run with $r > 0.00$ because type 3 agents are rigid in their behaviors: they always “choose” to contribute 20 MUs whatever the others did before. Since $r \cong 0.10$ in the runs that better approximate the empirical results, this implies that about 10% of agent will contribute their full endowment in the final period. Human beings, the strong altruists as well, are weaker disciples of Kant and usually reduce (though less than conditional cooperators) their contribution in response of others’ free-riding. In order to keep simple the model and explore the limits of the strong reciprocity argument, our agents are straightforward in their strategies and, consequently, rather rigid. More complex models, e.g. allowing wider possibilities of behavioral change even to strong reciprocators, may, of course, be designed. This would generate greater empirical consistence, but also would weaken the theoretical argument.

On the other hand, notwithstanding the limits presented above, our model reproduces the empirical results with a fair degree of accuracy. A first, fundamental finding is, therefore, that, by using player types corresponding to simple and strong reciprocity schemes besides rational egoist actors, it is possible to interpret and replicate the experimental findings of repeated PG

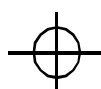




games. The parameter values $p \cong 0.50$, $q \cong 0.40$ and $r \cong 0.10$ that led to the runs closest to the empirical results are themselves significant. They sketch a world split, almost equally, between selfish and cooperative people: a picture which is probably not too far from reality, at least in situations where reputation and social control have little room. However, the main result is perhaps that rational egoists and simple reciprocators alone are not sufficient to obtain a good approximation of the punishment condition dynamics. A fundamental, though small, quota of strong reciprocators is necessary in order to achieve effective punishment and, consequently, in order to buster the rise in cooperation in the punishment condition. This is consistent with some recent studies arguing for strong reciprocity as a necessary (but not sufficient) schema in the interpretation of multilateral cooperation and altruistic punishment situations and, more generally, in the understanding of human altruism (Bowles & Gintis, 2004; Fehr & Fischbacher, 2003; Gintis, *et al.*, 2002). Those studies usually start from the fact that simple reciprocity schemes are insufficient in explaining the empirical findings of both field and experimental researches. Specifically, when cooperation and/or altruistic punishment (i.e. the punishment of subjects behaving against the collective interest) arise in situations when it is unlikely that the benefit conferred to others will be repaid either directly or indirectly, any analysis in terms of simple reciprocity can not hold. For instance, in one shot Ultimatum games, simple reciprocity can not account for the fact that most second players reject unfair offers (see Camerer, 2003:43-117 for a review of experimental results), while strong reciprocity is able to correctly predict that second players will sanction unfair first players by refusing their offers (Fehr & Fischbacher, 2003). The plausibility of the evolution of strong reciprocity has been also theoretically demonstrated by Samuel Bowles and Herbert Gintis (2004). Bowles and Gintis used agent-based simulation to reproduce the conditions of the hunter-gatherer group conditions of the Pleistocene. Their work shows how a mixed population, including a substantial share of strong reciprocators, may evolve starting from a population composed by selfish agents only and it convincingly advocates in favor of the plausibility of the strong reciprocity argument.

The results presented above are instead based on the comparison between simulated and real world environments. However, they advocates as well in favor of the idea of a model including multiple types of players (including some strong reciprocators) as an important schema in the understanding of human behavior. However, the fact that the model works in explaining the dynamics of PG games both with and without decentralized punishment institutions shows only its sufficiency, not its necessity.⁷ In order to increase its strength, future researches may follow a double strategy. First, the model should be applied to simulate the empirical outcomes of interaction schemes others than PD/PG situations. Second, new researches should aim both to directly identify rational egoist, simple reciprocator and strong reciprocator subjects in empirical settings and to test the consistence of their actual behaviors with the theory. The use of this double approach will permit to reach stronger evidence in a close future.

⁷ More generally, John Holland (1998, 241) argues that the rigorous definition of a computer-based model guarantees that the starting conditions of the model are sufficient for any phenomena observed during its execution, not that they are necessary.



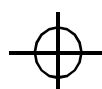


Acknowledgements

Authors thank Riccardo Boero, Silvana Dalmazone, Matteo Galizzi, Matteo Richiardi and Flaminio Squazzoni for their valuable comments.

References

- Ahn, T. K., E. Ostrom, D. Schmidt, J. Walker. (2003a), 'Trust in Two-Person Games: Game Structures and Linkages.' In E. Ostrom, & J. Walker (eds.). *Trust and Reciprocity*:323-351. New York: Russell Sage Foundation, .
- Ahn, T. K., E. Ostrom, & J. Walker. (2003b). 'Heterogeneous preferences and collective action.' *Public Choice*, 117: 295-314.
- Andreoni, J. (1995). 'Cooperation in Public-Goods Experiments: Kindness or Confusion?' *American Economic Review*, 85: 891-904.
- Bowles, S., & H. Gintis. (2004). 'The evolution of strong reciprocity: cooperation in heterogeneous populations.' *Theoretical Population Biology*, 65: 17-28.
- Burlando, R. M., & R. Guala. (2005). 'Heterogeneous Agents in Public Goods Experiments.' *Experimental Economics*, 8: 35-54.
- Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.
- Ledyard, J. (1995). 'Public Goods: A Survey of Experimental Research.' In Kagel, J., & A. Roth (eds.). *Handbook of Experimental Economics*:111-194. Princeton: Princeton University Press.
- Fehr, E., & U. Fischbacher. (2003). 'The nature of human altruism.' *Nature*, 425: 785-791.
- Fehr, E., & S. Gächter. (2000). 'Cooperation and Punishment in Public Goods Experiments.' *American Economic Review*, 90: 980-994.
- Fehr, E., & S. Gächter. (2002). 'Altruistic punishment in humans.' *Nature*, 415: 137-140.
- Fehr, E., & K.M. Schmidt. (1999). 'A theory of fairness, competition and cooperation.' *Quarterly Journal of Economics*, 114: 817-868.
- Fischbacher, U., S. Gächter, & E. Fehr. (2001). 'Are people conditionally cooperative? Evidence from a public goods experiment.' *Economics Letters*, 71: 397-404.
- Fowler, J. H., T. Johnson, & O. Smirnov. (2005). 'Egalitarian motive and altruistic punishment.' *Nature* 433, E1 (online version).
- Gintis, H. (2000). 'Strong Reciprocity and Human Sociality.' *Journal of Theoretical Biology*, 206: 169-179.
- Gintis, H., E. Alden Smith, & S. Bowles. (2001). 'Costly Signaling and Cooperation.' *Journal of Theoretical Biology*, 213: 103-119.
- Gintis, H., S. Bowles, R. Boyd, & E. Fehr. (2003). 'Explaining altruistic behavior in humans.' *Evolution and Human Behavior*, 24: 153-172.
- Hamilton, W. D. (1964). 'The Genetical Evolution of Social Behavior.' *Journal of Theoretical Biology*, 7: 1-52.
- Hammerstein, P., (ed.). (2003). *Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: The MIT Press.





- Holland, J. H. (1998). *Emergence: From Chaos to Order*. New York: Basic Books.
- Hayashi, N., E. Ostrom, J. Walker, & T. Yamagishi. (1999). 'Reciprocity, trust, and the sense of control: A cross-societal study.' *Rationality and Society*, 11: 27-46.
- Isaac, R. M., J. Walker, & A.W. Williams. (1994). 'Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups.' *Journal of Public Economics*, 54: 1-36.
- Kurzban, R., & D. Houser. (2005). 'Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations.' *Proceedings of the National Academy of Sciences*, 102:1803-1807.
- Ostrom, E. (2000). 'Collective Action and the Evolution of Social Norms.' *Journal of Economic Perspectives*, 14: 137-158.
- Ostrom, E. (2003). 'Toward a Behavioral Theory Linking Trust, Reciprocity, and Reputation.' In E. Ostrom, & J. Walker, (eds.). *Trust and Reciprocity. Interdisciplinary Lessons from Experimental Research*:19-79. New York: Russell Sage Foundation.
- Ostrom, E., & J. Walker, (eds.). (2003). *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*. New York: Russell Sage Foundation.
- Ostrom, E., J. Walker, & R. Gardner. (1992). 'Covenants With and Without a Sword: Self-Governance is Possible.' *American Political Science Review*, 86: 404-17.
- Richerson, P. J., R.Y. Boyd, & J. Henrich. (2003). 'Cultural Evolution of Human Cooperation.' In Hammerstein, P., (ed.). *Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: The MIT Press.
- Trivers, R. L. (1971). 'The Evolution of Reciprocal Altruism.' *Quarterly Review of Biology*, 46: 35-57.
- Wright, J. (1999). 'Altruism as a signal: Zahavi's alternative to kin selection and reciprocity.' *Journal of Avian Biology*, 30:108-115.
- Yamagishi, T. (1986). 'The Provision of a Sanctioning System as a Public Good.' *Journal of Personality and Social Psychology*, 51: 110-116.
- Zahavi, A., & A. Zahavi. (1977). 'The Handicap Principle: A Missing Piece of Darwin's Puzzle.' Oxford: Oxford University Press.

