

Altruism may arise from individual selection

Angel Sánchez and José A. Cuesta

Grupo Interdisciplinar de Sistemas Complejos (GISC)
Departamento de Matemáticas
Universidad Carlos III de Madrid
28911 Leganés, Madrid, Spain

May 20, 2004

Keywords Altruism, Strong Reciprocity, Individual Selection, Evolutionary Theories, Behavioral Evolution

The fact that humans cooperate with non-kin in large groups, or with people they will never meet again, is a long-standing evolutionary puzzle with profound implications.^{1,2} Cooperation is linked to altruism, the capacity to perform costly acts that confer benefits on others.³ Theoretical approaches had so far disregarded costly acts that do not yield future benefits for the altruist, either directly^{4,5} or indirectly.⁶⁻⁹ Recently, strong reciprocity,^{10,11} i.e., the predisposition to cooperate with others and to punish non-cooperators at personal cost, has been proposed as a schema for understanding altruism in humans.^{3,12} While behavioral experiments support the existence of strong reciprocity,^{11,13-15} its evolutionary origins remain unclear:¹⁶⁻¹⁸ group and cultural selection are generally invoked to compensate for the negative effects that reciprocity is assumed to have on individuals.¹⁹⁻²³ Here we show, by means of an agent-based model inspired on the Ultimatum Game,²⁴ that selection acting on individuals capable of other-regarding behavior can give rise to strong reciprocity. The results, consistent with the existence of neural correlates of fairness,²⁵ are in good agreement with observations on humans^{11,13-15} and monkeys.²⁶

Substantial evidence in favor of the existence of strong reciprocity comes from experiments using the so-called Ultimatum Game,²⁴ and from agent-based models^{21,23,27} (see Refs. 3, 12 for summaries). In the Ultimatum Game, under conditions of anonymity, two players are shown a sum of money, say 100 €. One of

the players, the “proposer”, is instructed to offer any amount, from 1 € to 100 €, to the other, the “responder”. The proposer can make only one offer, which the responder can accept or reject. If the offer is accepted, the money is shared accordingly; if rejected, both players receive nothing. Since the game is played only once (no repeated interactions) and anonymously (no reputation gain), a self-interested responder will accept any amount of money offered. Therefore, self-interested proposers will offer the minimum possible amount, 1 €, which will be accepted.

In actual Ultimatum Game experiments with human subjects, average offers do not even approximate the self-interested prediction. Generally speaking, proposers offer respondents very substantial amounts (50 % being a typical modal offer) and respondents frequently reject offers below 30 %. Most of the experiments have been carried out with university students in western countries, showing a large degree of individual variability but a striking uniformity between groups in average behavior. The fact that indirect reciprocity is excluded and that interactions are one-shot allows one to interpret rejections in terms of strong reciprocity.^{10,11} A large study in 15 small-scale societies¹³ found that, in all cases, respondents or proposers behave in a reciprocal manner. Furthermore, the behavioral variability across groups was much larger than previously observed: while mean offers in the case of university students are in the range 43%-48%, in the cross-cultural study they ranged from 26% to 58%.

In order to assess the possible evolutionary origins of these behaviors, we introduce and analyze here a drastically simplified model. Imagine a population of N players of the Ultimatum Game with a fixed sum of money M per game. Random pairs of players are chosen, of which one is the proposer and another one is the respondent. We will assume that players are capable of other-regarding behavior (empathy); consequently, in order to optimize their gain, proposers offer the minimum amount of money that they would accept. Every agent has her own, fixed acceptance threshold, $1 \leq t_i \leq M$ (t_i are always integer numbers for simplicity). Agents have only one strategy: respondents reject any offer smaller than their own acceptance threshold, and accept offers otherwise. Although we believe that this is the way in which ‘empathic’ agents will behave, in order not to hinder other strategies *a priori*, we have also considered the possibility that agents have two independent acceptance and offer thresholds. As we will see below, this does not change our main results and conclusions. Money shared as a consequence of accepted offers accumulates to the capital of each of the involved players. As our main aim is to study selection acting on modified descendants, hereafter we interpret this capital as ‘fitness’ (here used in a loose, Darwinian sense, not in the more restrictive one of reproductive rate). After s games, the agent with the overall minimum fitness is removed (randomly picked if there are several) and a new

agent is introduced by duplicating that with the maximum fitness, i.e., with the same threshold and the same fitness (again randomly picked if there are several). Mutation is introduced in the duplication process by allowing changes of ± 1 in the acceptance threshold of the newly generated player with probability $1/3$ each. Agents have no memory (i.e., interactions are one-shot) and no information about other agents (i.e., no reputation gains are possible).

Figure 1 shows that strong reciprocity, in the form of altruistic punishment, can be selected for at the individual level in small populations ranging from $N = 10$ to $N = 10\,000$ agents when selection is strong ($s = 1$). The initial distribution of thresholds rapidly leads to a peaked function, with the range of acceptance thresholds for the agents covering about a 10% of the available ones. The position of the peak (understood as the mean acceptance threshold) fluctuates during the length of the simulation, never reaching a stationary value for the durations we have explored. The width of the peak fluctuates as well, but in a much smaller scale than the position. At certain instants the distribution exhibits two peaks (see distribution at 7.5 million games). This is the mechanism by which the position of the peak moves around the possible acceptance thresholds. Importantly, the typical evolution we are describing does not depend on the initial condition. In particular, a population consisting solely of self-interested agents, i.e., all initial thresholds are set to $t_i = 1$, evolves in the same fashion. The value M of the capital at stake in every game is not important either, and increasing M only leads to a higher resolution of the threshold distribution function.

The success of reciprocators does not depend on the selection rate (although the detailed dynamics does). Figure 2 shows the result of a simulation with 1000 agents in which the removal-duplication process takes place once every $s = 10\,000$ games. To show further that the initial conditions are irrelevant, for this plot we have chosen an initial population of self-interested agents. As we may see, the evolution is now much less noisy, and the distribution is narrower, becoming highly peaked and immobile after a transient. The value of s at which this regime appears increases with the population size. The final mean acceptance threshold at which simulations stabilize depends on the specific run, but it is very generally a value between 40 and 50. We thus see that the selection rate may be responsible for the particulars of the simulation outcome, but it is not a key factor for the emergence of strong reciprocity in our model.

The behavior of our model has to be compared with the results of previous studies of the Ultimatum Game by Page and Nowak.^{28,29} The model introduced in those works has a dynamics completely different from ours: following standard evolutionary game theory, every player plays every other one in both roles (proponent and respondent), and afterwards players reproduce with probability pro-

portional to their payoff (which is fitness in the reproductive sense). Simulations and adaptive dynamics equations show then that the population ends up composed by players with fair (50%) thresholds. This is different from our observations, in which we hardly ever reach an equilibrium (only for large s) and even then equilibria set up at values different from the fair share. The reason for this difference is that the Page-Nowak model dynamics describes the $s \rightarrow \infty$ limit of our model, in which between death-reproduction events the time average gain all players obtain is the mean payoff with high accuracy. We thus see that our model is more general, in so far as we can study regimes far from the standard evolutionary game theory limit. As a result, we find a variability of outcomes for the acceptance threshold consistent with the observations in real human societies.^{3,12,13} Remarkably, another context in which this regime (finite s) is the relevant one is genetic algorithms, where in every evolution step only a small fraction of the population interacts.³⁰

To further confirm the differences between our approach and Page and Nowak's one, we have considered the same alternative as they did, namely to assign agents a new strategical variable, o_i , defined as the amount offered by player i when acting as proponent, and subject to the same mutation rules as the acceptance threshold, t_i . While Page and Nowak observed that in their setup, this modification of the model led to fully rational players (i.e., in our model, $t_i = o_i = 1$), except for fluctuations due to mutations. Figure 3 shows clearly that in our model the dynamics remains very complicated and equilibria are never reached within the duration of our simulations. Once again, this is due to the fact that the dynamics we propose does not remove the fluctuations of the payoff obtained by the players as the limit $s \rightarrow \infty$ does. In other words, in our model the effects of finite time between generations and of stochasticity play a non trivial role and sustain strong reciprocity (existence of players with $t_i > 1$) even if acceptance and offer obey independent rules. In connection with this, it is interesting to note that the interplay between randomness and finiteness of the population leads to changes in the view of the evolutionary stability of cooperation.³¹ This and our present report suggest that general approaches beyond standard evolutionary game theory may provide insights into the issue of how cooperation arises.

Evolutionary explanations of strong reciprocity have been advanced in terms of gene-culture coevolution.^{19-23,27} The underlying rationale is that altruistic behavior leads to fitness disadvantages at the individual level. But why must strong reciprocators have lower fitness than other members of their group? While alternative compensating factors (e.g., sexual selection) have been suggested,¹⁷ our results show clearly that, in the context of the Ultimatum Game, altruistic punishment¹⁴ may be established by individual selection alone. Our simulations are consistent

with the large degree of variability among individuals^{3,12} and among societies,¹³ and reproduce the fact that typical offers are much larger than self-interested ones, but lower than a fair share. While in our model agents have other-regarding behavior (empathy), i.e., agents offer the minimum they would accept if offered to them, this is not a requisite for the emergence of strong reciprocators as the two-threshold simulations show. The population evolves by descent with modification and individual selection, as the model does not implement cultural (other than parent-to-child transmission) or group selection of any kind. To be sure, we do not mean that these mechanisms are irrelevant for the appearance and shaping of altruism: what we are showing is that strong reciprocity (and hence altruism) may arise in their absence. Observations of strongly reciprocal behavior in capucin monkeys,²⁶ where cultural transmission, if any, is weak, strengthens this conclusion. Further support for our thesis comes from reports of individual, pre-existent acceptance thresholds shown by neural activity measurements in Ref. 25. In this respect, neural mechanisms gratifying cooperation as those demonstrated in Ref. 32 may have evolved to reinforce behaviors selected for at the individual level as we are suggesting. The detrimental effects of unfair sanctions on altruism¹⁵ is yet another piece of evidence in favor of the existence of such individual acceptance ('fairness') thresholds. Our conclusion that altruism does not necessarily have negative consequences for individuals draws such theories nearer to a biological perspective. Indeed, our results suggest that, despite its not being self-evident, altruistic strategies may do better in terms of fitness than selfish ones, even without repeated interactions or reputation gain. This conclusion, which would imply that strictly speaking there is no truly altruistic behavior, may have far-reaching implications in decision-making models and the design of public policies.^{17,18}

References

- [1] Darwin, C. *The Descent of Man, and Selection in Relation to Sex* (Murray, London, 1871).
- [2] Gould, S. J. *The Structure of Evolutionary Theory* (Harvard University Press, Cambridge, 2002).
- [3] Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791 (2003).
- [4] Trivers, R. L. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).

- [5] Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
- [6] Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
- [7] Milinski, M., Semmann, D. & Krambeck, H. J. Reputation helps solve the ‘tragedy of the commons’. *Nature* **415**, 424–426 (2002).
- [8] Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. Lond. B* **268**, 745–753 (2001).
- [9] Gintis H., Smith, A. E. & Bowles, S. Costly signaling and cooperation. *J. Theor. Biol.* **213**, 103–119 (2001).
- [10] Gintis, H. Strong reciprocity and human sociality. *J. Theor. Biol.* **206**, 169–179 (2000).
- [11] Fehr, E., Fischbacher, U. & Gächter. Strong reciprocity, human cooperation and the enforcement of social norms. *Hum. Nat.* **13**, 1–25 (2002).
- [12] Gintis, H., Bowles, S., Boyd, R. & Fehr, E. Explaining altruistic behavior in humans. *Evol. Hum. Behav.* **24**, 153–172 (2003).
- [13] Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. & McElreath, R. In search of *Homo Economicus*: Behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* **91**, 73–78 (2001).
- [14] Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
- [15] Fehr, E. & Rockenbach, B. Detrimental effects of sanctions on human altruism. *Nature* **422**, 137–140 (2003).
- [16] Hammerstein, P., ed. *Genetic and Cultural Evolution of Cooperation. Dahlem Workshop Report 90* (MIT Press, Cambridge, MA, 2003).
- [17] Bowles, S., Fehr, E. & Gintis, H. Strong reciprocity may evolve with or without group selection. *Theoretical Primatology*, December 2003.
- [18] Vogel, G. The evolution of the golden rule. *Science* **303**, 1128–1130 (2004).
- [19] Richerson, P. J., Boyd, R. T. & Henrich, J. Cultural evolution of human cooperation. In Ref. 16, pp. 357–388.

- [20] Henrich, J. & Boyd, R. Why people punish defectors. *J. Theor. Biol.* **208**, 79–89 (2001).
- [21] Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. USA* **100**, 3531–3535 (2003).
- [22] Gintis, H. The hitchhiker’s guide to altruism: Gene-culture co-evolution and the internalization of norms. *J. Theor. Biol.* **220**, 407–418 (2003).
- [23] Bowles, S., Choi, J.-K. & Hopfensitz, A. The co-evolution of individual behaviors and social institutions. *J. Theor. Biol.* **223**, 135–147 (2003).
- [24] Güth, W., Schmittberger R. & Schwarze, B. An experimental analysis of ultimate bargaining. *J. Econ. Behav. Org.* **3**, 367–388 (1982).
- [25] Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. The neural basis of economic decision-making in the ultimatum game. *Science* **300**, 1755–1758 (2003).
- [26] Brosnan, S. F. & de Waal, F. B. M. Monkeys reject unequal pay. *Nature* **425**, 297–299 (2003).
- [27] Bowles S. & Gintis, H. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theor. Popul. Biol.* **65**, 17–28 (2004).
- [28] Page, K. M. & Nowak, M. A. A generalized adaptive dynamics framework can describe the evolutionary ultimatum game. *J. Theor. Biol.* **209**, 173–179 (2000).
- [29] Page, K. M. & Nowak, M. A. Empathy leads to fairness. *Bull. Math. Biol.* **64**, 1101–1116 (2002).
- [30] Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, Reading, 1989).
- [31] Nowak, M. A., Sasaki, A., Taylor, C. & Fudenberg, D. Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650 (2004).
- [32] Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S. & Kilts, C. D. A neural basis for cooperation. *Neuron*, **35**, 395–405 (2002).

Acknowledgments This work owes much to group discussions at GISC, for which we thank its members, particularly Carlos Rascón for help with the literature. A.S. is thankful to Maxi San Miguel for introducing him to the subject. We acknowledge financial support from Ministerio de Ciencia y Tecnología (Spain).

Competing interests statement The authors declare they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.S (anxo@math.uc3m.es).

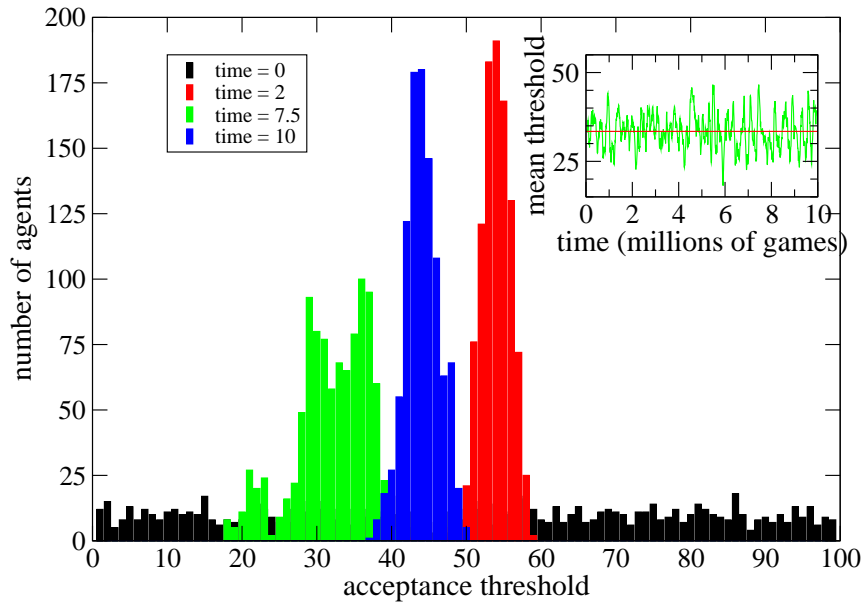


Figure 1: Non-self-interested behavior establishes spontaneously on small populations. Population size is $N = 1000$, the capital to be shared per game is $M = 100$. Death and birth takes place after every game ($s = 1$). Initial acceptance thresholds are distributed uniformly ($t_i = t_0$ conditions lead to the same output). Plotted are the distributions of acceptance threshold at the beginning of the simulation and after 2, 7.5 and 10 million games. Inset: Mean acceptance threshold as a function of simulated time, is averaged over intervals of 10000 games to reduce noise (in the raw data spikes appear that go above 50 or below 10). The red line in the inset is the average over all times of the mean, located at 33.45.

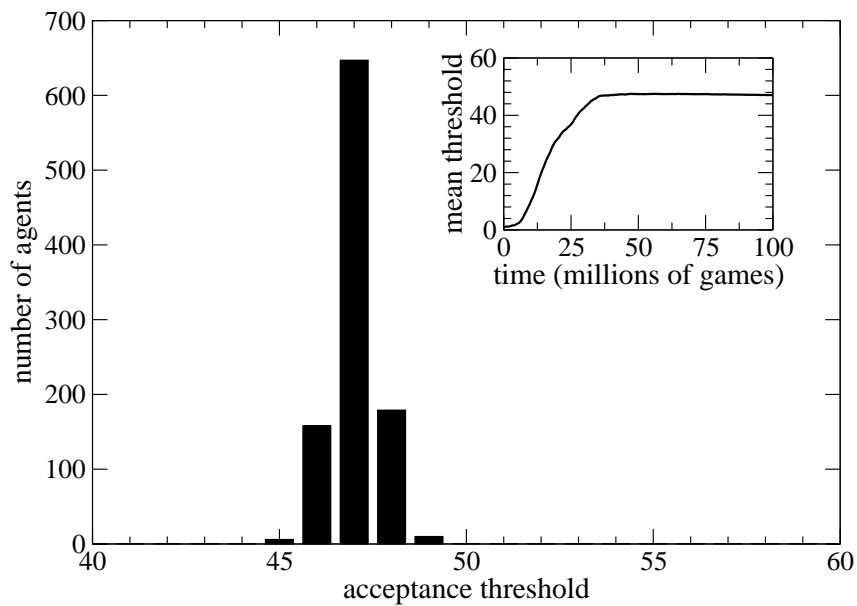


Figure 2: Slow selection rates lead to stationary acceptance threshold distributions very narrowly peaked. Population size is $N = 1000$, the capital to be shared per game is $M = 100$ and selection is weak ($s = 10\,000$). Initial agents are all self-interested ($t_i = 1$). Plotted is the distribution of acceptance threshold at the end of the simulation. There are no agents with thresholds outside the plotted range. Inset: Mean acceptance threshold as a function of simulated time. The asymptotically stable mean is very slowly approaching 47.

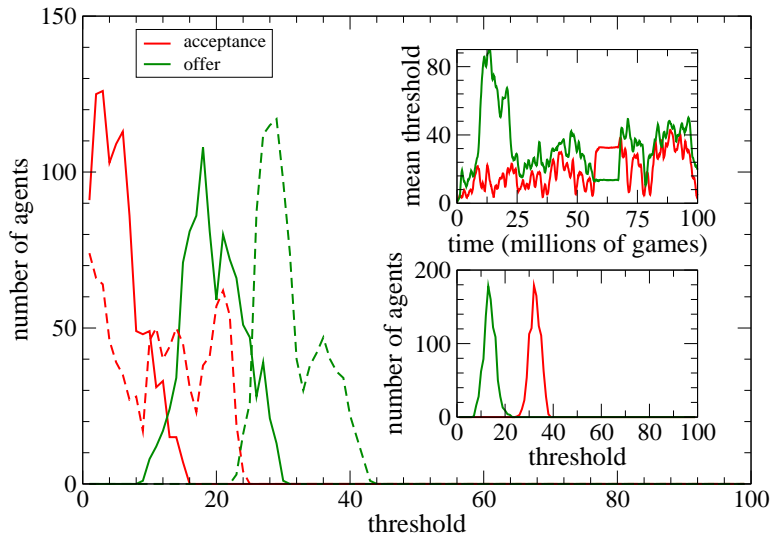


Figure 3: Introduction of an independent level for the amount of money offered by the agents does not change our conclusions. Population size is $N = 1000$, the capital to be shared per game is $M = 100$ and selection is intermediate ($s = 1000$). Initial agents are all fully rational ($t_i = o_i = 1$). Plotted are the distribution of acceptance threshold (red) and offered amount (green) after 50 million games (dashed) and 100 million games (solid). Upper inset: Mean acceptance threshold and offered amount as a function of simulated time. The offered amount is most of the time larger than the acceptance threshold, and occasional crosses lead to a very slow dynamics until the situation is restored (see the plateaus around 62.5 million games, and corresponding distributions in the lower inset).