

Altruism, altruistic punishment and social investment

Klaus Jaffe

Universidad Simón Bolívar (kjaffe@usb.ve), Caracas, Venezuela.

Abstract

The concept of altruism is used in very different forms by computer scientist, economists, philosophers, social scientists, psychologists and biologists. Yet, in order to be useful in social simulations, the concept “altruism” requires a more precise meaning. A quantitative formulation is proposed here, based on the cost/benefit analysis of the altruist and of society at large. This formulation is applied in the analysis of the social dynamic working of behaviors that have been called “altruistic punishments”, using the agent based computer model Sociodynamica. The simulations suggest that “altruistic punishment” by itself can not maintain altruistic behaviors. “Altruistic behavior” is sustainable in the long term only if these behaviors trigger synergetic forces in the society that eventually make them produce benefits to most individuals. The simulations suggest however that “altruistic punishment” may work as a “social investment”, and is thus better called “decentralized social punishment”. This behavior is very efficient in enforcing social norms. The efficiency of decentralized social punishment in enforcing norms was dependent on the type of labor structured of the virtual society. I conclude that what is called “altruistic punishment” emerges as a type of social investment that can evolve either through individual and/or group selection, as a successful devise for changing or enforcing norms in a society. Social simulations will help us in better understanding the underlying dynamic working of such devices.

Introduction

Altruism and reciprocity is theoretically unlikely to evolve as a result of natural selection. Yet such behaviors evidently exist, and not only among humans (Wilson 1976). No consensual explanation as to the forces allowing for its existence and maintenance exist (Sigmund et al. 2001, Johnson et al 2003). Thus human and animal cooperation and altruism continues to remain a puzzle. Among the most recent explanations proposed for solving this puzzle we might cite the following: Altruistic norms can ‘hitchhike’ on the general tendency of internal norms to be personally fitness-enhancing (Simon 1990) and that a multi-level selection, gene-culture co-evolution argument then explains why individually fitness-reducing internal norms are likely to be pro-social as opposed to socially harmful (Gintis 2003). Alternatively, neutral non-social players might stabilize the system enhancing the chances of altruistic behaviors to settle in social populations (Hauert et al. 2002). Another argument is based on the observation that reputation may foster social behavior among selfish agents, and is considerably more effective with punishment than with reward (Sigmund et al. 2001). Finally, I proposed that societies allow for the occurrence of synergies that provide hidden or retarded benefits to the various actors, allowing natural selection to maintain behaviors that cement social bonding, such as altruism (Jaffe 2001, 2002a). Evidence for long term hidden benefits of apparently altruistic behavior has been reported in the context of religions (Berman 2003).

A more recent argument put forward to explain the maintenance of altruistic and of reciprocal behaviors is altruistic punishment. Support that the punishment of non-cooperators at a cost to the punisher and the punished may favor adaptation of altruistic behavior has been provided by field studies (Fehr and Gächter 2002). Simulation experiments suggest that selection on small sized groups (Boyd and Richerson 1992) or asymmetries between altruistic cooperators and altruistic punishment, allows altruistic punishment to evolve in population (Boyd et al. 2003). Thus, some authors claim that altruistic punishment can sustain cooperation when altruistic participation cannot (see review in Gintis 2003). Yet, not all are convinced of this argument (Sigmund et al. 2001, Hauert et al. 2002, Johnson et al. 2003). Part of the contradiction may arise because of differing views on the subject, caused by different intuitive definitions of altruism, based on the traditions of the different disciplines involved. Most of this discussion, however, has been related to a problem in evolutionary biology related to the likelihood of group selection forces to work (see Sober and Wilson 1999 for example), and little effort has been made in actually demonstrating the existence of possible benefits to the group of altruism, a fact more of interest to economists. Group selection arguments do not solve the economic shortcomings of altruistic behavior (Jaffe 2002a), nor do they explain where the necessary synergy emerges that allows for the putative benefits of altruism (Jaffe 2001). Thus, here I want to reconcile the view of economists with that of biologists.

Looking for the term “altruism” in an economic dictionary (Pearce 1996), we find it defined as “Concern for the well-being of others. Thus, the utility function of an individual may exhibit the form: $U_i = U(X, Y, U_j)$ where X and Y are goods consumed by individuals i, and U_j is the utility of another individual, j.” A biology dictionary (Lincoln et al 2001) defines altruism as “The situation in which one individual acts to promote or enhance the fitness of an unrelated individual or of other members of a group at the same time reducing its own fitness: the case of mutually beneficial behavior is known as reciprocal altruism”. Wilson (1976) defines altruism simply as “Self-destructive behavior performed for the benefit of others”. Thus, two overlapping but distinct concepts are given the same name in different disciplines. Interestingly, most of the literature dealing with altruism lacks a reference to any of the definition of altruism just given or to any other. Researchers seem to rely on common sense when using the concept of altruism.

The original definition of “altruism” is traced to Auguste Comte, a French mathematician and philosopher during the first half of the 1800s. Comte, the founder of positivism, defined it as the antithesis of egoism but without giving it a specific definition by its own. (Comte 1830-1842). In general, altruism defines those acts that aim the good of others. The Encyclopedia Britannica refers to this definition as the “ethical” definition. The Encyclopedia Britannica defines altruism among animals as the performance of an act from which animals derive no direct benefit, and which often is very costly to the animals performing it.

Biological evolution is known to be able to allow the development of such behaviors if they favor close kin (Hamilton 1964), which thus will be able to transmit the genes, on which the altruistic behavior is based, to future generations (see Wilson 1976 for an extensive overview). Yet kin-selection does not explain many instances of altruistic behavior (see review in Jimenez-Alonso 1998, for example). Thus, other explanations, besides kin-selection theory, have been proposed to explain the behaviors we observe in nature. One such proposition supposes the existence of interactions that produce synergies, which benefit in the long term the individuals participating in the interactions (Jaffe 2001). These behaviors, if viewed in the short term, might be described as

purely altruistic and detrimental to the altruist, but have a high adaptive value, even under strict individual selection scenarios, when viewed using larger time windows.

The biological definition assumes that altruism is a behavior that benefits society at a cost to the altruist. This aspect, although not mentioned explicitly in many publications, is very often accepted as intuitively true. This aspect however contradicts the economic definition of altruism which assumes that altruism increases the utility function of individuals and thus increases also the aggregate utility function of society, leading to the perception that society would be better off, in aggregate economic terms, if altruism was more widely practiced. Jaffe (2002a) reported that 97.6% of the scientists related to economy, computer simulations and mathematical biology interviewed agree with this believe, without any experimental proof for it. In the same paper, the logical inconsistencies between both definitions were revealed, using computer simulations showing that altruism without synergistic economic benefits actually reduce the benefits of society in aggregate terms.

These definitions, when applied to concrete situations, produce even more contradictory interpretations. Specifically, the ethical definition of altruism, when describing the behaviors of animals or of agents in artificial societies, is utterly unpractical. The ethical definition of altruism is based on the use the aims of a behavior as a criterion for its description. As we know from the dynamics of complex systems, the aim or original intention of behaviors, and what they achieve in practice, are often very discordant. This phenomenon is colloquially referred to as “the law of the unexpected consequences”. Thus, in order to be useful in social simulations, the concept “altruism” requires a precise instrumental mathematical definition that can be applied to human social behavior as well as to animal or artificial societies. In order to achieve this, I will simulate a specific behavior as the frame for this exercise.

Social simulations are one way to try to disentangling these differing views. The simulation of artificial societies is a powerful tool that allows us to bridge the gap between different sciences. One such gap is represented by the meaning of the concept “altruism”. Very different concepts are behind references to “altruism” by computer scientists, economists, philosophers, social scientists, psychologist and biologists. In order to be useful in social simulations, the concept altruism, thus, requires a more precise mathematical definition.

The agent based simulation Sociodynamica (Jaffe, 2002a) provides a framework that allows testing the robustness of mathematically defined behaviors related to the various types of cooperation and “altruistic” interactions in which social organisms engage. When trying to answer fundamental questions, such as “Does society benefits from altruistic acts?” the simulations using the model Sociodynamica revealed a somewhat counterintuitive result. They showed that for most situations that have loosely been called altruistic, this behavior does not benefit the population, but rather, if put in pure economic terms, lower society's level of aggregate wealth. “Altruistic” behavior favored society at large only in the case of synergistic altruism and/or when special phenomena increased the indirect benefits of the altruistic act non-linearly; making the benefits obtained by the recipient(s) of the altruistic act greater than the costs incurred by the altruist. Extrapolating the results of these simulations to real live forces us to suggest that, in many situations that have been described as “altruistic”, non-economic factors, or factors not accounted for by the observer, such as: future security, access to privileges, prestige, loving care, transcendence, etc., may provide hidden benefits to the “altruist”.

Redefining “altruism”

The above reported result, that in most situations were “altruists” donate wealth to less wealthy agents, the aggregate wealth of society decreases, harming societies total well being, seems contradictory. “Altruism” that does not benefit society (even if we believe that it does) can hardly be called “altruism”. In order to avoid this kind of misrepresentation of reality, I propose here to formalize the definition of “true altruism” as follows.

Any behavior benefits society in aggregate terms, expressed here as total benefit to society S , only if:

$$K < \int_t (A + B) \quad \text{and} \quad S = \int_t (A + B) - K > 0 \quad \dots\dots 1$$

Where K is the cost to the altruist, A is the benefit to the recipient(s) and B is the eventual benefit to the altruist, all in a long term perspective, so that these benefits have to be integrated over time.

The more demanding definition of biologists for altruism will require that in addition to equation 1, altruism occurs only when the costs are greater than the eventual benefits the individual may receive, so that the compounded benefit to the individual I has to be negative:

$$K > \int_t B \quad \text{and} \quad I = \int_t B - K < 0 \quad \dots\dots 2$$

Under this formulation, behaviors that benefit the actor, in addition to benefiting others, so that $K < \int_t B$ or $I > 0$ cannot be considered altruistic and are better classified as enlightened egoistic acts that benefit the *polis* or community. I propose to call this type of behaviors “social investment”, as these behaviors would be equivalent to a Pareto investment that benefits eventually both the actor and society.

True altruism has to benefit society and therefore the condition $S > 0$ must hold for all altruistic acts. Thus, behaviors, in order to be called altruistic in a sense that satisfies both biologists and ethicists, have to meet conditions 1 and 2 simultaneously. Equations 1 and 2 can be complied with simultaneously only if $A > K$. This condition was referred to as “synergistic altruism” (Jaffe 2002a). Here, the benefit that the altruistic act conveyed to the receiver was greater than the cost to the altruist in producing the behavior; or the utility achieved by the receptor of the donation was larger than the loss in utility incurred by the donor.

In many known behaviors, $S < 0$ and $I < 0$, which achieves dissipation of wealth. These cases correspond to destructive behaviors that dissipate utility and are not likely to be evolved through biological evolution. Most of the behaviors intuitively defined as altruistic fall into this category (Jaffe 2002a). They can not be considered as adaptive behaviors to biologists. Behaviors that do not comply with equation 1, so that $S < 0$, can be called “destructive egoism” or just “destructive behavior”. Examples are parasitism, exploitation and destructive competition. Summarizing the possibilities defined by equations 1 and 2, we have:

$S > 0$ and $I > 0$: “social investment”
 $S < 0$ and $I > 0$: “destructive egoism”
 $S > 0$ and $I < 0$: “true altruism”
 $S < 0$ and $I < 0$: “destructive behavior”

For many behaviors described in the literature as “altruistic”, the apparent costs to the altruist agent is greater than his expected benefits only in the short term. That is

$K > B$ in the short term; whereas $K < B$, in the longer term.

This is often the case when we overlook non-direct economic benefits in calculating the long term utility of the putative altruist, such as future security, access to privileges, prestige, insurance against future liabilities, etc. In this case, it is hard to speak of “altruism”. If classical economic theory is applied, the phenomenon could be referred to as a long term investment. The central theoretical achievement of classical and neoclassical economics, is based on these kinds of investments. The ensuing dynamics is summed up by Adam Smith's (1776) metaphor of the 'invisible hand', that the interaction of selfish economic agents may produce a mutually beneficial and Pareto-optimal outcome. This outcome can be shown to be a robust equilibrium outcome using simulations with the model Sociodynamica (Jaffe 2002a).

A recently described form of cooperation between human subjects has been called “altruistic punishment”. Altruistic punishment means that individuals punish transgressors of fair play, although the punishment is costly for them and yields no direct material gain. Fehr & Gächter (2002) proposed that altruistic punishment might explain the fact that people frequently cooperate with genetically unrelated strangers, often in large groups, with people they will never meet again, and when reputation gains are small or absent. These patterns of cooperation cannot be explained by the nepotistic motives associated with the evolutionary theory of kin selection and the selfish motives associated with signaling theory or the theory of reciprocal altruism. Fehr & Gächter (2002) showed experimentally that the altruistic punishment of defectors is a key motive for the explanation of cooperation. Using human subjects, they showed that cooperation flourishes if altruistic punishment is possible, and breaks down if it is ruled out. They also showed evidence indicating that negative emotions towards defectors are the proximate mechanism behind altruistic punishment, confirming previous findings (Rawlings 1968).

In this paper I use the agent based simulation model Sociodynamica to explore the limits and meaningfulness of the concepts defined above, and show that the behavior called “altruistic punishment” complies with equation 1 but not necessarily with equation 2, suggesting that this family of behaviors is not necessarily altruistic and is better described as a decentralized social investment.

Methods

The agent based computer simulation Sociodynamica was used to study the effect of altruistic punishment on aggregate wealth accumulation in artificial societies. A somewhat simpler version of Sociodynamica was published before (Jaffe 2002a, c). The model simulates a continuous two-dimensional toroidal world through which different types of agents wandered with Brownian motion, each at its proper speed. The speed of this motion (m) ranged from 0-30 pixels / time step.

Agents could not learn. The simulations tested for the survival abilities of agents under variable circumstances. As dead agents were substituted by new ones, which had their parameters assigned at random, the simulations served as a way of weeding out those combination of parameters that conferred low fitness or low survival capabilities to agents, selecting those agents possessing parameters that conferred them larger survival possibilities. Agents did not inherit their parameters, as Sociodynamica is a metaphor for a society of agents living in a free competitive market.

The toroidal world was supplied with patches of agricultural land (food resources: R_f) and mines (mineral resources: R_m). Each time an agent happened to land over one of these resources while walking randomly around, they acquired a single unit (w_o) of the corresponding resource, accumulating wealth, either as food (w_f) and/or as mineral wealth (w_m).

Agents spend some of their wealth in food in order to survive, consuming food at a basal constant rate (b), which was a fraction of the resource unit (w_o). The wealth in food (w_f) of each agent changed each time step:

$$dw_f = -b + w_o \quad \text{where } w_o = 0 \quad \text{if no resources are encountered.}$$

b determined the degree of external constraints or of competitiveness of the environment and was fixed at 0.1, indicating the speed of degradation of accumulated resources in w_o / time-step. This value produced simulation outcomes that are closed to what we expect in real societies (Jaffe 2002a,c). Agents with no food resources left ($w_f = 0$) perished and were substituted by a new agent with randomly assigned parameters. This substitution process allowed maintaining the total number of agents in the population constant.

Similarly, agents encountering minerals acquired a single unit of the resource (w_o) each time they encountered it. Minerals never degraded ($b_m = 0$). The wealth in minerals (w_m) was inversely related to the probability of sudden death for each agent. That is, mineral wealth improved the odds of surviving external constraints. External “catastrophes” killed agents at random, each time step, and large amounts of w_m protected the agents against these catastrophes by reducing the probability of being affected by them. Agents with $w_m = 0$ could survive, though, with a lower probability. The agents were struck by a fatal catastrophe if the following relation was true:

$$w_m < \text{rnd}(0-1) * D$$

So that the greater the wealth of accumulated minerals of the agent, the lower their probability of being struck by a catastrophe, at any level of danger (D)

Agents moved in random directions each time step. Each time an agent met another at a distance smaller than 20 pixels, an exchange of wealth could occur. These could be of various types. Donations of food occurred when the difference in food wealth ($w_{f1} - w_{f2}$) between the two agents was larger than 2. Then the richer agent transferred food to the less wealthy. The amount of food transferred depended on the **generosity** (g) of the donating agent, which varied initially among agents from 0 to 5 deciles of their wealth (w_f), i.e. 0 to 50 % of their wealth.

Both types of resources were replenished continuously. Each of them was concentrated in a different single patch and the total amount of resources was 200 w_o for food and 100 w_o for

minerals. Each resource patch was distributed initially at random in the landscape but remained in the same place during the duration of each run.

In some simulations, more “**structured societies**” were simulated by modeling labor specialization of the agents. In this case, agents were subdivided into three categories. Farmers which specialized in collecting only food; miners which collected only minerals; and traders. Traders specialized in trading minerals for food when encountering a farmer, and food for minerals when encountering a miner. Traders increased the value of minerals (w_m) they traded by 50 %. This increase in wealth simulated an “addition” of value of minerals due to “processing” or the effect of “work productivity” (see also Jaffe 2002c). When not explicitly stated, artificial societies had no structure, i.e. no division of labor, and all agents could collect food and/or minerals. No traders were simulated in non-structured societies.

The parameters that were explored in the simulations presented in the paper were:

- 1- The amount of altruistic punishment dispensed, which was regulated by the amount of agents acting as **altruistic punishers** in the population. This amount was kept constant during each simulation run. The altruistic punisher reduced the wealth of the agent it encountered, if the degree of generosity of the encountered agent was $g \leq 1$ (agents that gave less than 10 % of their wealth when encountering poor agents, see above), at a cost to its own wealth.
- 2- The reaction of agents when punished. This was coded as **sensitivity to punishment**, which regulated the degree to which punished agents increased their generosity after receiving a punishment. This increase could vary from 0 to 5 deciles of w_f .

In the simulations presented here, the cost to the punisher was fixed at 10 % of its w_f , and equaled the cost the punished agent paid. The values of parameters not specifically analyzed or described as being fixed in a simulation, were allowed to vary randomly among agents in the ranges mentioned above. Unless stated otherwise, simulations were run 200 times with 500 agents for 100 time steps. Although the configuration of populations never stabilized completely, after 60 time steps changes were very small. Thus, the populations of agents reached a stable state after time interval of 100 time steps.

Results

Any combination of variables can be run using the model Sociodynamica that is available on the web (<http://atta.labb.usb.ve/Klaus/klaus.htm>). Simulations exploring the effect of synergistic and other interactions are presented in Jaffe 2002a. Here, I will present only a selected choice of simulations, summarized as follows.

Figure 1 shows the frequency distribution of two sets of 1500 agents with regard to the degree of generosity shown by each agent after 100 time steps. In these simulations, agents with low generosity ($g \leq 1$) were punished (Altruistic punishment), or were not punished (No punishment). Sensitivity to punishment was set to one, indicating that agents increased in one decile of their w_f the amount of wealth they donated, after receiving the punishment. Punishment of non-generous individuals increased the frequency of generous individuals in the population slightly and changed the frequency distribution of generosity among agents very significantly. The

equilibrium frequency distribution of generosity among surviving agents subjected to altruistic punishment was divergent; agents either profited from not being generous at all, or were sufficiently generous so as to avoid punishment.

Non-linear behavior was also evidenced when studying the effect of the “sensitivity to punishment” on the average generosity of the population (Figure 2). The simulations showed that even with a low sensitivity to punishment (i.e. agents increase their level of generosity after punishment only by a single decile of wf), the effect on the aggregate generosity was very significant.

The intensity of altruistic punishment, expressed as the percent of altruistic punishers in the population, affected the average level of generosity of agents also non-linearly (Figure 3). The effect of the percentage of altruistic punishers on the level of average generosity was relatively small. Much greater was the effect of the percentage of altruistic punishers on aggregate wealth achieved by this artificial society (Figure 4). If this effect was measured in simulations using more structured societies (i.e., societies where agents could be farmers, miners or traders), then the population showed a remarkable increase in the susceptibility of its level of generosity to the percentage of altruistic punishers present (altruistic punishers could be farmer, miners or traders). Even a very small percentage of altruistic punishers affected the level of generosity expressed by the whole population (Figure 5). On the other hand, the effect of the percentage of altruistic punishers on aggregate wealth (Figure 6) was relatively lower than in the case of non-structured societies shown in Figure 4. Note that in non-structured societies, the aggregate wealth achieved was much higher than that of structured societies (compare absolute values in y-axis of Figures 4 and 6).

Yet interestingly, as can be seen from Figures 4 and 6, the societies performed best at the aggregate level if no altruism was exercised, which occurred in the case where no altruistic punishers were included in the simulation (0 % of altruistic punishers). As the amount of altruistic punishers increased, increasing the altruist in the population, the aggregate wealth accumulated decreased. Clearly, altruistic punishment did not overcome the negative effects of altruism on society in this case.

The dynamic behavior of structured societies differs from that of non-structured societies in various ways. The average level of generosity achieved by the agents in structured societies was more susceptible to variations in the percentage of altruistic punishers present (Figure 5), compared to societies with no division of labor among agents (Figure 3). On the other hand, the average generosity reached by agents in societies with no labor structure responded stronger to the degree of sensitivity to punishment simulated (Figure 2), compared with that of structured societies (Figure 9). In societies with no labor structure, agents reached older ages (Figure 7), compared to simulations with structured societies (Figure 8). This last result reflects the fact that mortality of agents in societies with labor structure was higher, as agents had a lower probability of having the right placement in space and the right motility for the type of labor they had to perform, compared to agents in societies with no labor structure. In societies with no division of labor, the type of labor performed by the agent was a consequence of the type of placement in space and the type of motility the agent possessed.

The susceptibility of structured societies to the percentage of altruistic punishers present (Figure 9), correlated with the number of agents exercising the three different tasks (Figure 10). Figure 10 presents the number of farmers, miners and traders present in the society when dynamic equilibrium was reached after 100 time steps, for simulations with different susceptibilities to punishment. As the sensitivity to punishment increased, the number of farmers increased whereas that of miners and traders decreased. The differences in average generosity observed between simulations were due mainly to changes in the number of agents showing no generosity (Figure 11). This suggests that in more structured societies, it is more difficult for individuals to avoid compliance with social norms as the agent's survival is more dependent on society

Figure 1: Frequency of occurrence of agents in an artificial society showing various levels of generosity, in simulation where up to 30 % of agents could engage in altruistic punishment, or in simulations with no punishment.

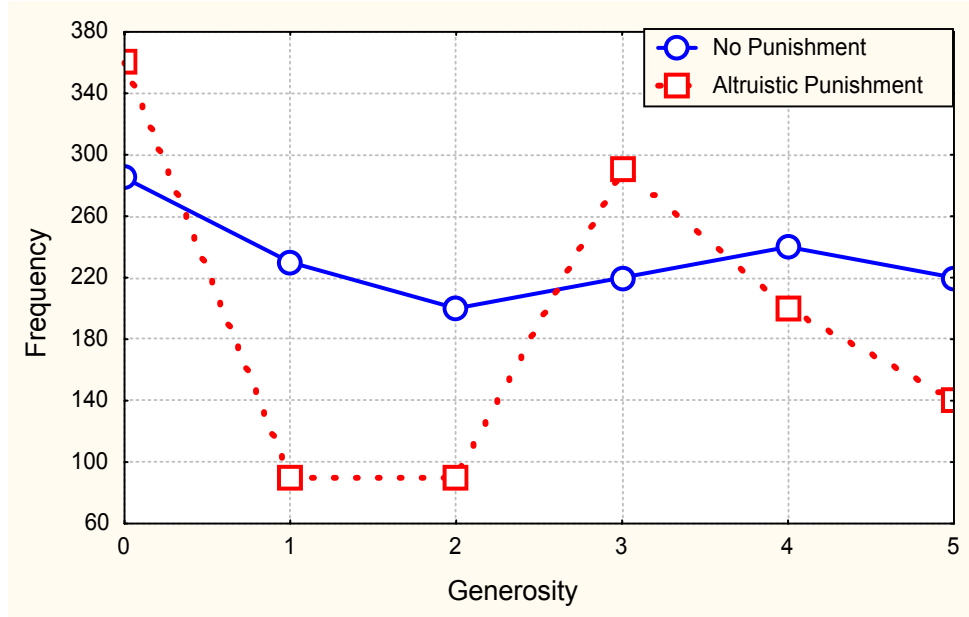


Figure 2: The average level of generosity shown by 1500 agents, after 100 time steps, in simulations with different levels of susceptibility to punishment, i.e. different levels of generosity increase by agents receiving a punishment.

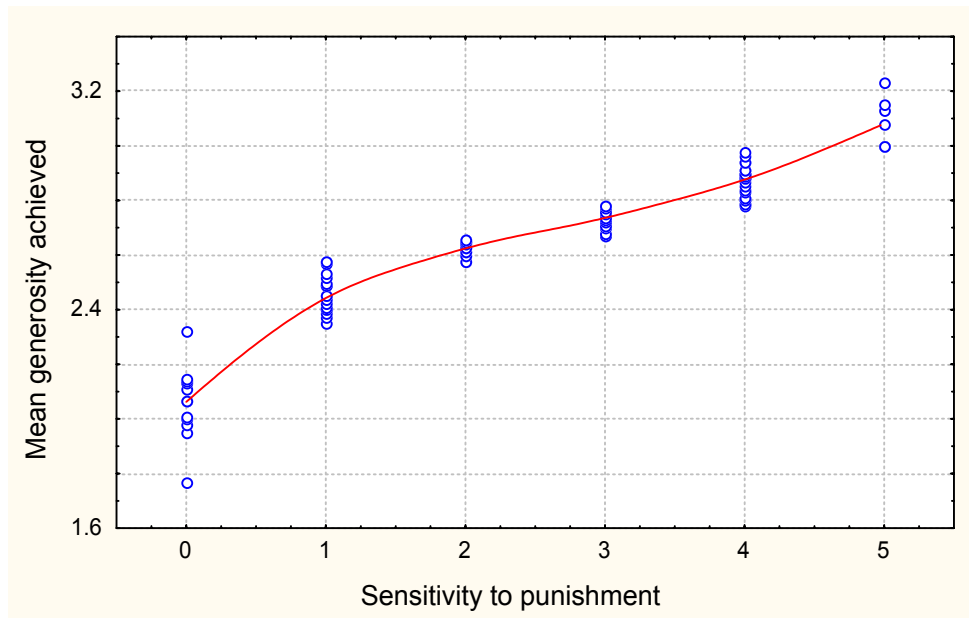


Figure 3: Average level of generosity shown by 500 agents in simulations having different percentages of agents exercising altruistic punishment.

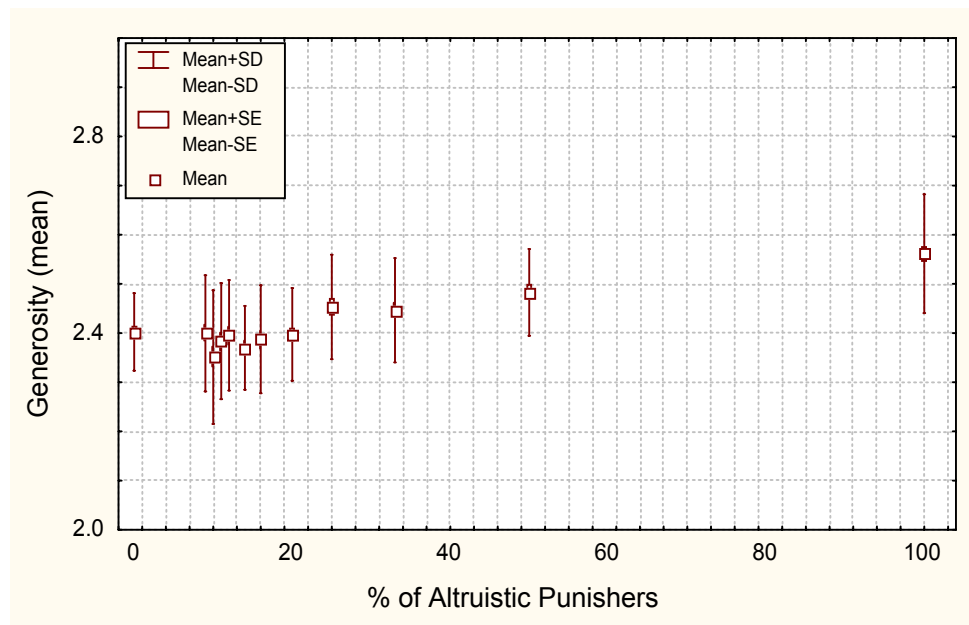


Figure 4: Aggregate total wealth (GDP) achieved by the population of 500 agents in simulations having different percentages of agents exercising altruistic punishment.

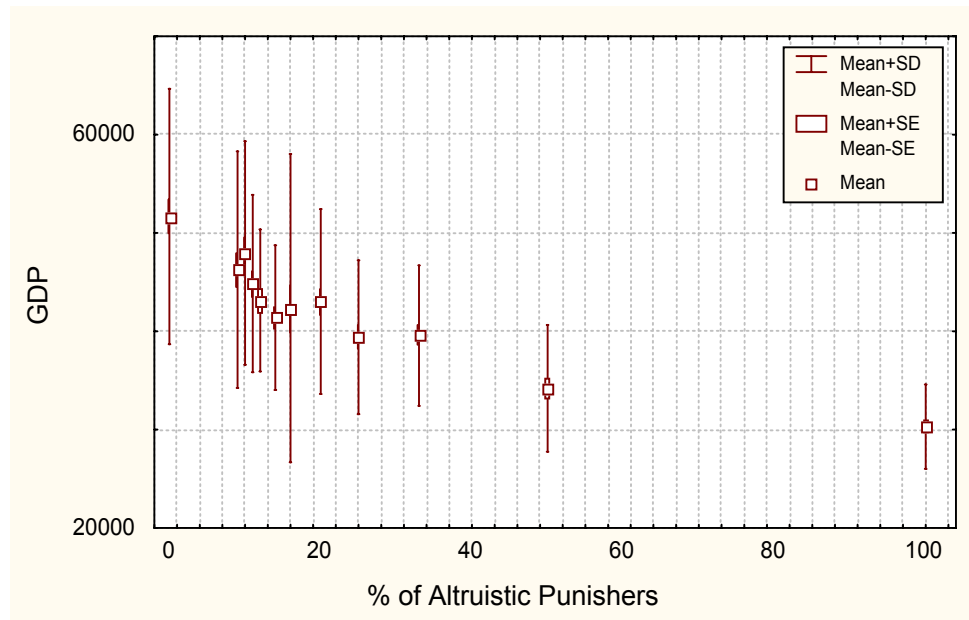


Figure 5: Average level of generosity shown by 500 agents in simulations having different percentages of agents exercising altruistic punishment, when agents divided labor (farmers, miners, traders).

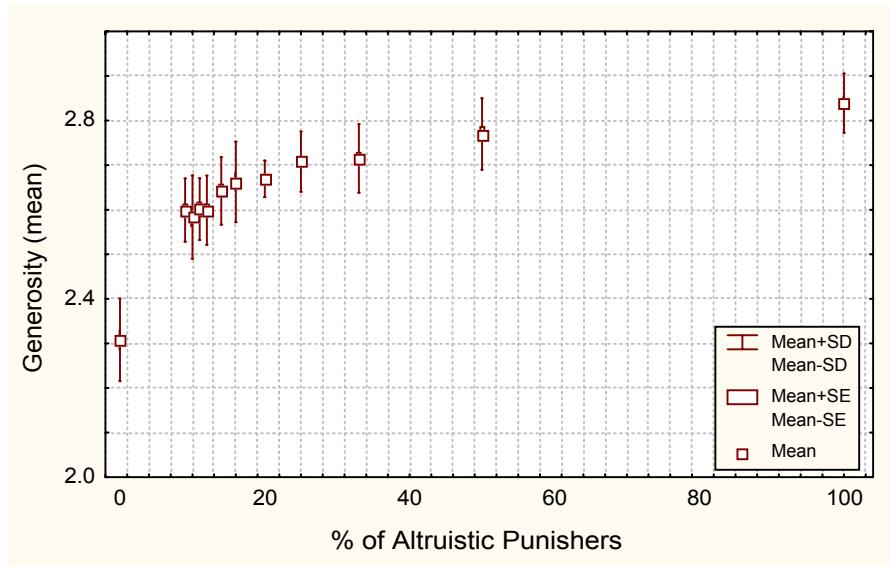


Figure 6: Aggregate total wealth (GDP) achieved by the population of 500 agents in simulations having different percentages of agents exercising altruistic punishment, when agents divided labor (farmers, miners, traders).

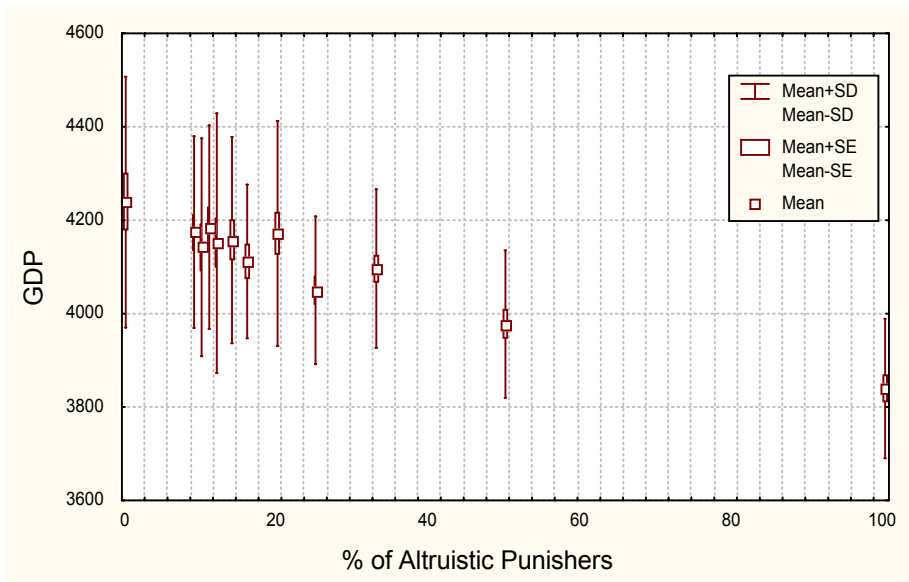


Figure 7: Age achieved by 500 agents after 100 time steps in simulations with different levels of susceptibility to punishment, when agents had no division of labor.

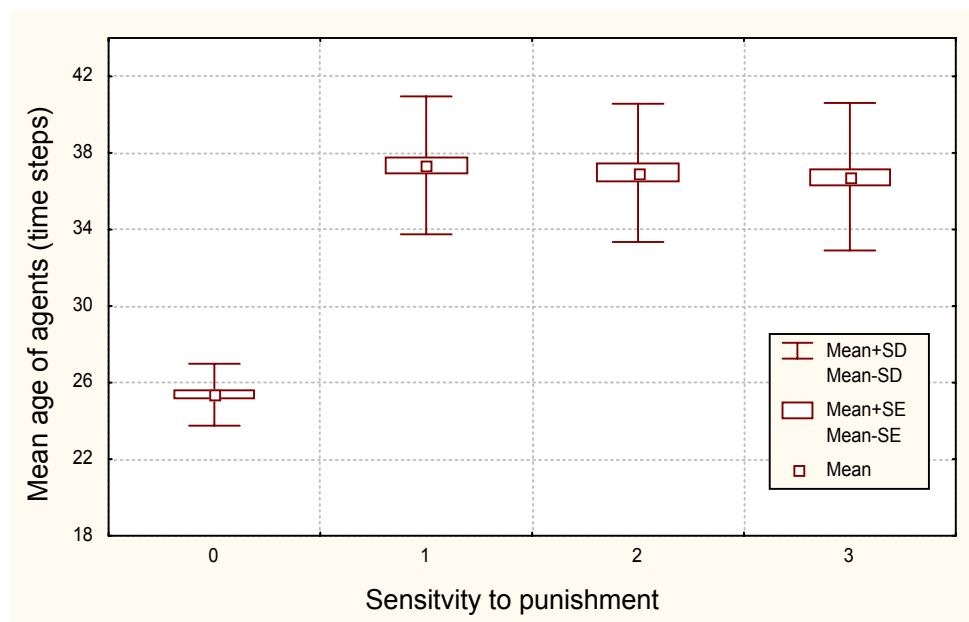


Figure 8: Age achieved by 500 agents after 100 time steps in simulations with different levels of susceptibility to punishment, when agents had division of labor (farmers, miners and traders).

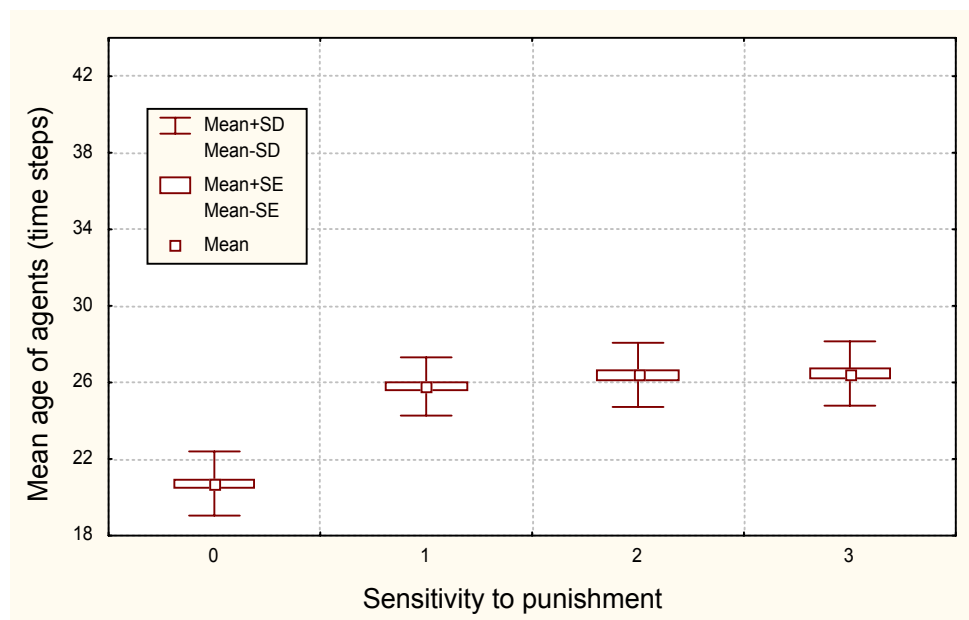


Figure 9: The average level of generosity shown by 500 agents, after 100 time steps, in simulations with different levels of susceptibility to punishment, when agents had division of labor (farmers, miners and traders).

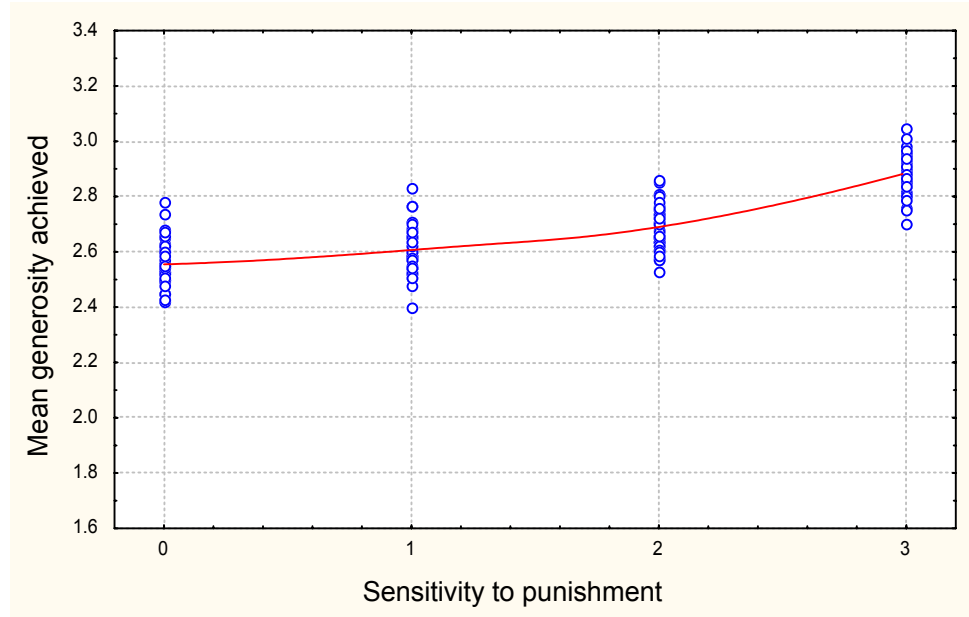


Figure 10: Percentage of agents in artificial societies, after 100 time steps, in simulations shown in Figure 9. The bars and whiskers represent the mean and standard deviation of 200 simulations

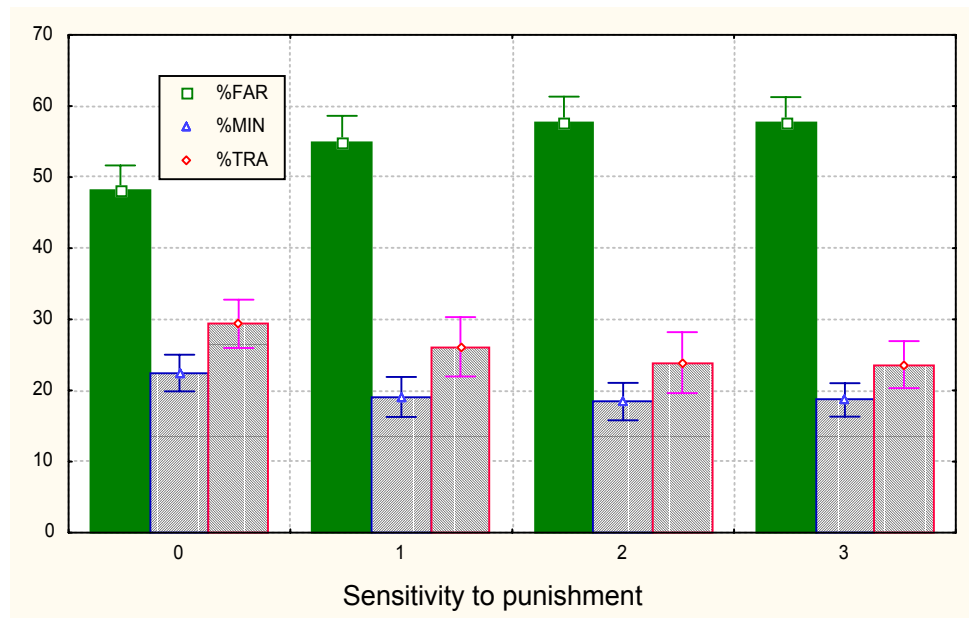
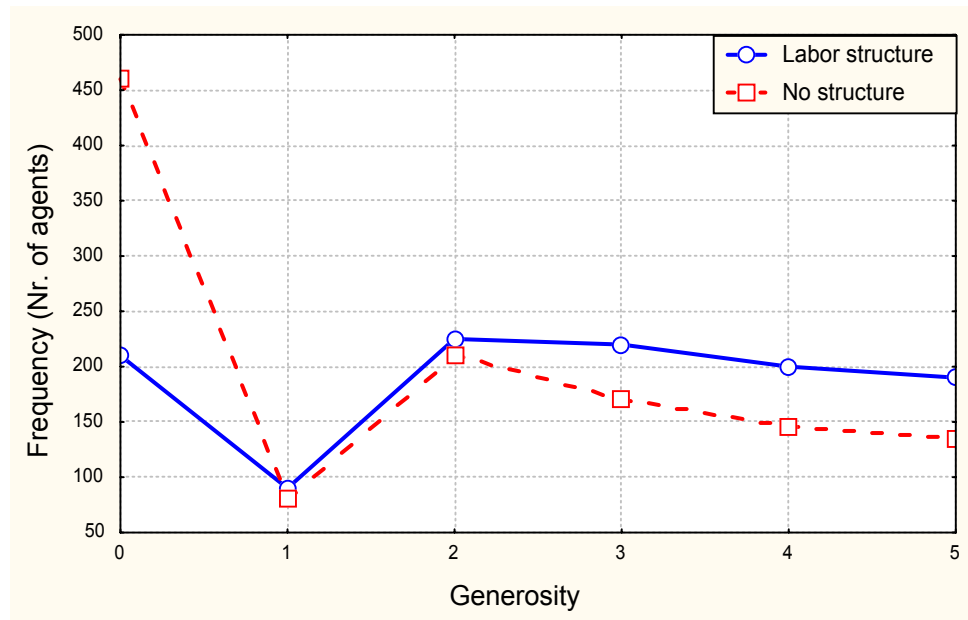


Figure 11: Frequency of occurrence of agents in an artificial society showing various levels of generosity, in simulation where 50 % of agents engaged in altruistic punishment.



Discussion

The main result of the simulation is that altruism as defined here will not be sustainable in populations, small or large, even if altruistic punishment is practiced. This lack of adaptive potential is evident at both, the individual and the group level (here represented at the aggregate level). Altruism, even of the altruistic punishment kind, will be sustainable only if the altruistic behavior triggers some synergistic forces in society that increase the fitness of the members of the population (Jaffe 2001), or increases the return to the utility function of the actors (Jaffe 2002a). In the case of altruistic punishment, it will be sustainable in time only if it helps to enforce behaviors that will be beneficial to society. Thus, it is the profitability of the behaviors enforced, or the social effect of the rules enforced by altruistic punishment that will determine their adaptive value or their resilience in cultural evolution.

Yet, the simulation results presented here showed that punishment, altruistic or otherwise, is a strong and efficient driving force in societies for the enforcement of social norms.

The definition of altruism given here aims at providing a precise benchmark for assessing altruism. It is based on the assumption that an altruistic act is costly to the individual and beneficial to society ($S > 0$ and $I < 0$). Other definitions of altruism exist, but the one proposed here, can be clearly defined in mathematical terms and is relevant for both, the study of biological evolution and socio-economic interactions. Yet, better names for the various phenomena studied here may appear in the future. For example, many charities try to fuse the meanings of altruism with that of social investment. Their aim is clearly not an analytical understanding of both concepts, but their

insistence in mixing both terms hints to the closeness of their meaning. The concept of “social investment”, as used here ($S > 0$ and $I > 0$), does not imply that agents are rational. They could be bounded rational or totally irrational and still perform social investments. Thus, the term does not reflect intentions but outcomes of behaviors. I hope that a future meeting on interdisciplinary nomenclature may propose better names for the concepts discussed here. But even if my verbal definitions may draw criticisms, their mathematical definitions are unambiguous, simple and applicable to any discipline.

In all the situations explored here, K (the cost to the “altruist”) was considered to be a one-time cost. Of course, K might be conceived as a long term cost that might even vary in time. This last assumption will introduce very interesting dynamic effects into the model and should be explored in the future.

Even as the simulation results showed that an increase in altruistic generosity, if not accompanied by a synergistic effect, depresses the aggregate wealth accumulated by society (see also Jaffe 2002a); and that this effect has as a consequence that altruistic punishment *per se* is not beneficial to society; the simulations showed that altruistic punishment is very effective in changing the behavior of a society. In the simulations presented here, the change of behavior achieved by altruistic punishment was the degree of generosity shown by richer agents towards poorer ones. If the behavioral change achieved through altruistic punishment is beneficial to society, then altruistic punishment will have a beneficial effect on society. We might thus improve equation 1 by proposing the following equation:

$$K < \int_t (A + B + C)$$

where K is the cost to the donor or actor of the altruistic act, A is the benefit to the recipient or recipients, B is the benefit of the donor and C the benefit for third parties, in a long term perspective.

In order for altruistic punishment to be called unequivocally altruistic, it has to comply with equation 1 and 2 simultaneously. With the improved formulation of equation 1, this can be achieved if third parties benefit from the behavior that is costly to the altruist. It is however more probable in real life situations that all actors, the punisher, the punished and third parties may benefit in the long term from the increased compliance of the social norm enforced by “altruistic punishment”, making the behavior, in the terminology defined above, a “social investment”. In this light, the class of behaviors described by Fehr & Gächter (2002), seem better referred to as “Decentralized Social Punishment”, in the sense of decentralized social control defined in multiagent simulations (see Castelfranchi 2000 or Kaminka & Tambe 2000). The fact that this class of behaviors may benefit both, the individual exercising them and the group, make them evolvable by adaptive forces (in the biological sense) in both individual and group selection scenarios.

Another situation where altruism has been invoked is indirect reciprocity by image scoring (Nowak & Sigmund 1998). Also in this case, the observed behavior may have alternative explanations to true altruism (Jaffe 2002b) and may be described as a social investment. In general, in order to improve cross disciplinary communication, the use of the term altruism should be avoided, unless it is clearly defined and unless experimental evidence supports the assumptions of its definition. Many authors, although not explaining it explicitly, assume that altruism occurs when

an individual acts so as to benefit others rather than himself. In a social context, though, such behaviors very often are egoistic in the long term, as the benefit the actor behests others may reach him. The main difference of the present proposition to former uses of the concept “altruism” is that here, the dichotomy between altruistic and egoistic is not of “self” and “others”, but of “individual” and “society”. The proposed definition is more objective (no need to define a “self”) and thus potentially more useful for researcher in social simulations and other interdisciplinary sciences. Thus, I propose to call behaviors “altruistic” only if it can be shown that they comply with equations 1 and 2.

Decentralized social punishments, as defined above, can be viewed as a social investment that helps to enforce social norms. Examples from real life might include the “whistle blower” denouncing corruption in corporations and institutions, the community watcher, and the honorary informal teacher. More institutional or centralized figures of social punishment might be as, or even more, efficient than decentralized social punishment. Such figures may include the constabulary and other types of police. Yet, it is the decentralized version, that was revealed by Fehr & Gächter (2002), that seems especially interesting, and field studies, aiming to quantify the occurrence of such behaviors, might help us in better understanding the working of our societies.

An interesting finding of the simulation results is the fact that labor specialization increases the susceptibility of the system to social coercion mechanisms such as “punishment”. That is, “social investments” work better the more structured the society. This feature justifies *a posteriori* the use of the term “social investment”. As already noted by Adam Smith (1776), increased labor specialization may increase the efficiency of the system, and makes it also more vulnerable to social control. In the simulations presented here, this effect was due a stronger selection pressure on agents in more structured societies and due to the fact that in societies with labor structure, the fate of individuals is more closely linked with that of others, compared to societies with no labor structure. The fact the model was able to capture differences in the dynamics of social investment depending on the social structure simulated, suggests that social simulations may reveal more details of the working of altruistic behaviors in different societies. Animal behaviors underplaying the principle of decentralized social punishments, such as a sense of social justice, have been reported for monkeys (Brosnan & de Waal, 2003). A testable prediction that might be suggested from the present results is that the dynamics of social investment among monkeys and humans should differ. The relationship between the effects of “social investments” and the type of labor structure of a given society certainly merits further studies, with both, artificial and real societies.

References:

- Berman, E. 2003. Hamas, Taliban and the Jewish Underground: An Economist's View of Radical Religious Militias. NBR Working Paper 10004 <http://www.nber.org/papers/w10004>
- Boyd, R. and Richerson, P.J. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* **13**: 171-195.
- Boyd, R., Gintis, H., Bowles, S., Richerson, P.J. 2003. The evolution of altruistic punishment. *Proceedings of the National Academy of Science* **100**: 3531-3535.
- Brosnan, S.F. and de Waal, F.B.M. 2003. Monkeys reject unequal pay. *Nature* **425**: 297-299.
- Castelfranchi, C. 2000. Engineering Social Order 1. ESA working paper <http://lia.deis.unibo.it/confs/ESAW00/pdf/ESAW04.pdf>
- Compte, A. 1830-1842. Cours de Philosophie Positive. Hermann Ed, Paris. 1998.
- Fehr, E and Gächter, S. 2002. Altruistic punishment in humans, *Nature* **415**: 137-140
- Gintis, H. 2003. The hitchhiker's guide to altruism: gene-culture coevolution, and the internalization of norms, *Journal of Theoretical Biology* **220**: 407-418.
- Hamilton, W. D. 1964. The genetic evolution of social behavior. *Journal of Theoretical Biology* **7**: 17-18
- Hauert, Ch., De Monte, S., Hofbauer J. and Sigmund, K. 2002. Volunteering as Red Queen Mechanism for Cooperation in Public Goods Game, *Science* **296**, 1129-1132
- Jaffe, K. 2001. On the relative importance of Haplo-Diploidy, Assortative Mating and Social Synergy on the Evolutionary Emergence of Social Behavior. *Acta Biotheoretica* **49**: 29-42.
- Jaffe, K. 2002a. An economic analysis of altruism: Who benefits from altruistic acts? *Journal of Artificial Societies and Social Simulations* **5**: 3 <http://jasss.soc.surrey.ac.uk/5/3/3.html>
- Jaffe, K. 2002b. On sex, mate selection and evolution: an exploration. *Comments on Theoretical Biology* **7**: 91-107, 2002.
- Jaffe, K. 2002c. Monte Carlo exploration of mechanisms for the creation of aggregate wealth. Proceedings of IAREP/SABE Conference, Turku, Finland. <http://atta.labb.usb.ve/Klaus/MonteCarlo%20Explo%20of%20Wealth.htm>
- Jimenez-Alonso W. 1998. The role of kin selection theory on the explanation of biological altruism: a critical review. *Journal of Comparative Biology* **3**: 1-14.
- Johnson, D.D.P, Stopka, P. and Knights, S. 2003: The puzzle of human cooperation. *Nature* **421**, 911 - 912

Kaminka, G.A. and Tambe, M., 2000. Robust Agent Teams via Socially-Attentive Monitoring. [Journal of Artificial Intelligence Research 12](#): 105-147.

Lincoln, R., Boxshall, G. and Clark, P. 2001. A Dictionary of Ecology, Evolution and Systematics. Second edition. Cambridge University Press. 361pp.

Nowak, M.A. and Sigmund, K. 1998. Evolution of indirect reciprocity by image scoring. *Nature* **393**: 573

Pearce, D.W. (ed) 1996. The MIT Dictionary of Modern Economics. Forth edition. MIT Press, Cambridge Massachusetts. 474 pp

Rawlings, E.I. 1968. Witnessing harm to other: a reassessment of the role of guilt in altruistic behavior. *Journal of Personality and Social Psychology* **10**: 377-80

Sigmund, K., Hauert, Ch. and Nowak, M. A. 2001. Reward and punishment , *Proceedings of the National Academy of Science* **98**, 10757-10762.

Simon, H. 1990. A Mechanism for Social Selection and Successful Altruism, *Science* **250**:1665–1668.

Smith, A. 1776. An Inquiry about the Nature and Causes of the Wealth of Nations. The Library of Economics and Liberty. <http://www.econlib.org/library/Smith/smWN.html>

Sober, E. and Wilson, D.S. 1999. Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press, 700pp

Wilson, E.O. 1976. Sociobiology: A New Synthesis, Harvard University Press.