FACOLTA' DI ECONOMIA
UNIVERSITA' DI BOLOGNA
SEDE DI FORLI

Corso di laurea in Economia
delle Imprese Cooperative
e delle Organizzazioni Non profit

Endogenous Social Preferences, Heterogeneity and Cooperation

Luca Zarri

Working Paper n.51
Giugno 2008

in collaborazione con

AICCON
CULTURA COOPERAZIONE | NONPROFIT

Luca Zarri,
Università di Verona

Informazioni

Facoltà di Economia – Corso di Laurea in Economia delle Imprese Cooperative e delle ONP
Tel. 0543 374673 - Fax 0543 374660 – e mail: segreteria.ecofo@unibo.it
Web: **www.ecofo.unibo.it**

June 24, 2008

Michele Biavati[♣], Marco Sandri[♦] and Luca Zarri[♠]

Abstract

We set up an analytical framework focusing on the problem of interaction over time when economic agents are characterized by various types of distributional social preferences. We develop an evolutionary approach in which individual preferences are endogenous and account for the evolution of cooperation when all the players are initially entirely selfish. In particular, within motivationally heterogeneous agents embedded in a social network, we adopt a variant of the indirect evolutionary approach, where material payoffs play a critical role, and assume that a co-evolutionary process occurs in which subjective preferences gradually evolve due to a key mechanism involving behavioral choices, relational intensity and degree of social openness. The simulations we carried out led to strongly consistent results with regard to the evolution of player types, the dynamics of material payoffs, the creation of significant interpersonal relationships among agents and the frequency of cooperation. In the long run, cooperation turns out to be the strategic choice that obtains the best performances, in terms of material payoffs, and 'nice guys', far from finishing last, succeed in coming out ahead.

JEL Classification: B41; C73; D74; Z13.

Keywords: Behavioral Economics; Cooperation; Prisoner's Dilemma; Social Evolution; Heterogeneous Social Preferences; Indirect Evolutionary Approach

[♣] ITAM (Instituto Tecnologico Autonomo de Mexico)
[♦] University of Brescia
[♠] University of Verona

> "To a large extent, experiences and events of the past causally affect the choices made by human individuals. Rational choice theory, while assigning expectations of the future to their proper role in models of social phenomena, must somehow incorporate the fact that *choice behavior is not exclusively drawn from the 'front' but is also pushed from the 'rear'*" (Gueth and Kliemt, 1998; p. 378 – italics added)

1. Introduction

The evolution of cooperation has been extensively studied in the last decades in the social sciences. Well-known explanatory mechanisms now include reputation formation, kin selection (Hamilton, 1964), direct reciprocity (Trivers, 1971; Axelrod, 1984) and indirect reciprocity and costly signalling (Nowak and Sigmund, 1998; Gintis et al., 2001). However, despite the relevance of the above explanations, the emergence and sustainability of cooperation is still a major puzzle, insofar as we refer to large-scale societies in which agents are genetically unrelated and interaction is decentralized and increasingly anonymous. How can cooperation be endogenously enforced and be stable over time within such social environments, when clear material incentives to act selfishly exist and exogenous enforcement mechanisms are unavailable? With regard to these contexts, recent approaches identify solutions such as the possibility of non-participation (see Kurzban and Houser, 2004, among others) and the presence of an implicit social contract prescribing cooperation on the part of (homogeneous) self-interested individuals (Benchekroun and Long, 2007). Janssen (2006) explains cooperation among genetically unrelated people by referring to players' ability to recognize untrustworthy opponents. He uses simulation experiments with artificial agents and shows that evolution of cooperation can occur when agents are able to learn to recognize the trustworthiness of other individuals.

In order to provide a satisfactory answer to the 'puzzle of cooperation', a different line of thought is increasingly gaining ground at both theoretical and empirical level and today a large body of experimental evidence suggests a broadening of the traditional framework of *homo oeconomicus* conveyed by the rational choice paradigm and centred about the established category of *material self-interest* (see e.g. Mueller, 1986). The role of this crucial assumption in economic theory has aroused extensive debate again and again in the history of the discipline as well as among philosophers interested in the methodological foundations of economic science. In the last decades, a lot of economic experiments have persuasively shown that the so called 'selfishness axiom' turns out to be often violated by subjects' actual behavior. If we followed the basic rational choice theory prescriptions, admitting that universal egoism is the rule, how could we then explain why non-opportunistic behaviors often survive, both in the lab and in relevant real-life domains? Why do people give to charities? Why are individuals often willing to incur costs in order to reward kindness and punish unkindness? What is the reason why cooperative actions often take place in interactive environments within which no monetary incentives to do so are at work?

In an attempt to give account of such growing body of evidence about seemingly 'irrational' behaviors, a new explanatory road has been taken, laying stress on the complexity of human nature and the variety of human goals and motives. The endeavor is to show that there is no founded reason to confine economics exclusively to agents pursuing strictly material satisfactions leaving out of consideration any kind of moral imperatives and restraints. As Sen (1987) clearly remarked, "Universal selfishness as actuality may well be false, but universal selfishness as a requirement of rationality is patently absurd". In particular, as far as experimental studies on cooperation are

concerned, behavioral economists' main contribution has been to show that a lot of behavior observed in the lab is compatible with the idea that subjects are driven by so called *'social preferences'* (see Fehr and Gaechter, 2000 and Camerer, 2003), that is individual preferences with a social content capturing "the phenomena that people seem to care about certain 'social' goals, such as the well-being of other individuals or a 'fair' allocation among members in society, in addition to their own benefits" (Li, 2006; p. 1).

However, with regard to nonselfish behavior, a key question remains largely unanswered, up to now: where do so called 'social preferences' come from? How can we evolutionarily justify their emergence, within significant social interaction scenarios? In other words, it is important to understand whether it is true or not that selective incentives actually play against individuals who pay attention to other people's well-being and that therefore these motivational types should sooner or later disappear in any specific economic or social environment in which they occasionally play a role. In this regard, unlike what many theorists claim by endorsing such 'naïve evolutionary view', we defend the alternative thesis that selective incentives need not always favor self-interested individuals at the expenses of nonselfish ones. In particular, we claim that analyzing the implications of *other-regarding preferences* seems is one of the most promising routes along the path indicated above. Notwithstanding this, it is crucial, in addressing such an issue, to go beyond the simple identification of altruism with some versions of what has been framed as 'enlightened egoism', as if it were theoretically necessary to assume that some unconscious motive (say, self-esteem or status-seeking motivations) or genetic force (a 'selfish gene') is at work in driving individual behavior. These behavioral patterns are of course plausible and interesting in themselves, but it would be questionable to identify such individualistic motivational set as exhaustive with respect to the problem of the existence of other-regarding behavior. We must admit, at least in principle, that not all seemingly altruistic behavior underlies egoistic motivations, i.e. that the desire to help others can be nonstrategically motivated and inspire costly voluntary behavior without any expectation of present or future material rewards (see on this Fehr and Gaechter, 2000; Camerer, 2003).

In this paper, we focus on a population where players are (potentially) driven by different types of social preferences and address the issue of cooperation by providing a solution based on a novel enforcement mechanism: a *co-evolutionary* process involving subjective preferences and behavioral choices, crucially depending on the material consequences of social interaction, along the lines of the so called 'Indirect Evolutionary Approach' (IEA; see on this Gueth and Yaari, 1992). Our major goal is to see whether the above mechanism turns out to be a powerful device for the emergence of cooperative behaviors among (initially selfish) strangers[1]. In particular, we assume that players are driven by *distributional social preferences*, that is preferences over final payoff allocations, but that a significant degree of player type heterogeneity exists. Several studies suggest that heterogeneity as to player types is a promising direction to shed light on the issue of cooperation. Erlei (2006) presents a simple model based on heterogeneous other-regarding preferences that turns out to have a high predictive accuracy, with regard to experimental evidence. Rotemberg (2007) shows that heterogeneous social preferences can account for the experimental results of ultimatum and dictator games. We carry out a simulatory analysis on a heterogeneous population in which different 'experimentally focal' types are simultaneously considered, and show that, despite the presence of relevant material incentives to defect, cooperation is a stable medium-run outcome[2]. Our major finding is that the analytical combination of endogenous sociality and heterogeneity provides us with a novel account for the emergence of cooperation when material

[1] Like Janssen (2006), we set up a simulatory experiment and find conditions under which a population initially dominated by selfish individuals choosing to defect evolves towards a population in which both a high level of cooperation and other-regarding preferences prevail.

[2] For a recent paper where cooperation results from a trade-off between material incentives and individual values, see Tabellini (2008).

incentives to free ride are present and exogenous enforcement devices are not available. Further, we obtain analogous results within a social environment in which material payoffs play a critical role in making preferences endogenous and assortative matchings occur, with the probability of meeting a given player depending on the degree of relational intensity characterizing such interaction.

The structure of the remainder of the paper is as follows. Section 2 presents the analytical model. Section 3 contains the major results of the different simulations we carried out**.** Section 4 concludes.

2. The structure of the game

We explore the emergence and sustainability of cooperation among (initially fully) selfish players, within a large-scale population where agents are randomly matched in pairs[3]. As far as pairwise interactions are concerned, the basic game-theoretic framework is given by the established material Prisoner's Dilemma (PD) setting[4]. If we consider the infinitely repeated PD, Rubinstein (1986), by studying two-person supergames where each player is restricted to carry out his strategies by finite automata, makes clear that cooperation cannot be the outcome of a solution to this game. However, Axelrod (1984) succeeds in providing precise conditions for this to happen and, unlike several contributions on the theme, he manages to do this without altering the very nature of the interaction problem[5]. Similarly, Oltra and Schenk (1996) investigate the evolution of cooperation in a simulation model where (selfish) agents play a one-shot Prisoner's Dilemma against their neighbors. Here they show that cooperation can persist and spread over in the long run insofar as agents choose their strategies according to imitation rules and the neighborhood structures they are located in overlap. Unlike these interesting contributions, in the basic version of the model we attempt to go beyond the 'dilemma' without altering players' strategy set or neighborhood structure, by focusing instead on the role of motivational and relational factors, within an evolutionary framework: in particular, our purpose is to set up a model where, at individual level, agents are (i) potentially nonselfish and (ii) allowed to gradually *strengthen* or *weaken* mutual relationships, as time unfolds. Our major goal is to study such analytical structure with reference to a motivationally complex population composed of heterogeneous agents (see Section 2.2.), in order to investigate whether and under what conditions cooperation may emerge as a stable outcome within such a broader and more realistic scenario.

Let us consider the following bimatrix, containing players' material payoffs in correspondence of the four possible outcomes of each pairwise encounter:

		Player 2	
Player 1		C	D
	C	2 , 2	0 , 3
	D	3 , 0	1 , 1

[3] Frameworks in which trading environments are populated by a large number of agents who meet randomly have been recently studied, among others, by Camera and Casari (2007).

[4] In the one-shot material PD, the Pareto-efficient solution, that is Mutual Cooperation (MC), is unable to endogenously emerge insofar as the two agents are assumed to be driven by classic selfish preferences. Cooperation here is a strictly dominated strategy: at the individual level, it is in the interest of each agent to defect, independently of the opponent's strategy. This well-known 'unpleasant' result holds even if the game is repeated a finite number of times, insofar as agents are egoists and we apply the standard 'backward induction' reasoning (see Luce and Raiffa, 1956).

[5] In particular, by referring to the well-known 'tit for tat' strategy, he shows that in a world of egoists in which no exogenous central authority exists and where an initial cluster of agents rely on reciprocity, MC may be able to emerge.

Table 1. Material Prisoner's Dilemma

The game continues for a finite number of periods. Hence, the only Subgame Perfect Nash equilibrium is Mutual Defection (MD).

The first distinctive feature of the model is given by the following mechanism, linking *behavioral* and *relational* dimension at individual level: we assume that, as a consequence of repeated interaction, the interpersonal relationship between two specific players, say A and B, develops over time and can be formally captured by means of the 'degree of relational intensity' IR, which evolves in a discrete way on the basis of individuals' strategic choices. In particular, IR_{AB} *increases* (resp., decreases) in period t if A observes that in the previous stage of the game (t-1) player B made the *same* (a different) strategic choice as her, i.e. whenever A observes that either MC or MD occurred at t-1. The rationale behind the introduction of such a mechanism has to do with the substantial evidence available today with regard to the existence of a specific mechanism connecting the perceived *similarity* of behavioral choices with the degree of *empathy* emerging at interpersonal level[6]. Neuroscientific evidence increasingly shows that empathy is a key source of pro-sociality (see Singer and Fehr, 2005). As Adam Smith (1759; pp. 9-10) authoritatively observed: "When we see a stroke aimed and just ready to fall upon the leg or arm of another person, we naturally shrink and draw back our own leg or own arm; and when it does fall, we feel it in some measure, and are hurt by it as well as the sufferer".

2.1. The dynamics of cooperation with other-regarding preferences: empathic altruism

In order to make clear how the mechanism illustrated above works, let us start with the simple case involving interactions between two motivationally homogeneous players driven by genuinely other-regarding preferences. Hence, let us suppose that A and B's (symmetric) utility functions are:

$$U_A = (1-w_{AB})*\Pi_A + w_{AB}*\Pi_B \quad (1)$$

$$U_B = (1-w_{BA})*\Pi_B + w_{BA}*\Pi_A \quad (2)$$

where the parameter of sociality w denotes the degree of altruism of each individual ($0 \leq w < 1$). In this case, both A and B can be said to be 'partial altruists', as "between the frozen pole of egoism and the tropical expanse of utilitarianism (there is)… the position of one for whom in a calm moment his neighbour's utility compared with his own neither counts for nothing, nor "counts for one", but counts for a fraction" (Edgeworth, 1881) [7]. However, far from adopting a static view of altruism, we believe that it is important to develop a 'relational' perspective, that is to analyze

[6] In this regard, Hoffman (1995) remarks that, in the context of Krebs' experiments, "Subjects who were told that they were *similar to the other*, as compared to those told that they were *dissimilar*, gave more pronounced physiological responses when the other appeared to be experiencing pleasure or pain, reported that they *identified* more with the other, felt worse while the other waited to receive an electric shock, *and were more likely to behave altruistically toward the other.* This finding has special interest since, as noted earlier, the physiological response to another's distress appears to have a large involuntary component" (p. 22 – italics added).

[7] For a similar formalization, see Ledyard (1995). Levine's (1998) specification of altruistic preferences coincides with ours when he assumes that players have the same regard for altruistic and spiteful opponents. Janssen (2006), in his simulation experiment, assumes that players are driven by *social-welfare preferences*: they always prefer more for themselves and the other player (like partial altruists), but are more in favor of getting payoffs for themselves when they are behind than when they are ahead.

other-regarding preferences within a dynamic framework which driving their evolution over time: with reference to the extensive debate mentioned in Section 1, it is of interest to verify whether and under what conditions non-standard preferences are sufficient to bring about the emergence of stable cooperative solutions which enable 'nice guys', far from finishing last, to come out ahead. Hence, we assume that the degree of altruism is not a fixed stock but depends, in turn, on the level of interpersonal relational intensity IR. This dependence is captured by the following functional form:

$$w_A = IR_{AB} / (IR_{AB} + b) \qquad (\text{where } b \in \Re_+;\ f' > 0 \text{ e } f'' < 0) \qquad (3)$$

The different values of the parameter b bring about different possible time paths, provided that we assume a similar behavior of the players in all the periods[8]:

T	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
IR	0	1	2	3	4	5	6	7	8	9	10
w (b=3)	0	0,25	0,4	0,5	0,57	0,62	0,67	0,7	0,72	0,75	0,77
w (b=9)	0	0,1	0,18	0,25	0,3	0,35	0,4	0,44	0,47	0,5	0,52

Table 2. Time Path of IR and w

Further, we assume that in each period each agent chooses between C and D on the basis of a simple expected utility calculation and believes, in the light of adaptive expectations, that his opponent makes, at time t, the same strategic choice chosen at time t - 1. We chose to adopt this assumption, which works *against* the emergence of cooperation, in order to avoid any speculation of the players on the opponent's future behavior. That means that a cooperation's strategic choice cannot be instrumentally used in order to influence opponent's relational intensities and, therefore, the sociality parameter.

When the game starts (t = 0), since the agents are supposed not to know one another, we assume that $IR_{AB}=IR_{BA}=0$, which implies that $w_{AB} = w_{BA} = 0$. The related utility functions are:

$$U_A = \Pi_A$$

$$U_B = \Pi_B$$

Therefore, at t = 0 both players defect[9]. However, in the light of their (potentially) altruistic utility functions, at t = 1 both players will start assigning a positive weight to their opponent's payoff, as in t = 1, IR = 1. In order to find out the exact stage of the game where the strategic change occurs, it is sufficient to calculate the value of w such that, say, for A the (expected) utility associated with cooperation (C) turns out to be greater than the level of utility generated by defection (D):

[8] Our formalization implies that altruism depends positively on relational intensity, but after a certain number of periods this takes place in a less than proportional way.

[9] At t = 0, it is as if both A and B, who do not know each other, were standard selfish agents. Hence, they both defect as defection is the dominant strategy in the standard one-shot PD setting.

$U_A(A_c / B_d) = (1-w_{AB})*0 + w_{AB}*3$

$U_A(A_d / B_d) = (1-w_{BA})*1 + w_{BA}*1$

Therefore, A will change his strategic choice when $U_A(A_c / B_d) > U_A(A_d / B_d)$, that is when

$3*w > 1 \quad \Rightarrow \quad w > 1/3$

and, substituting this critical threshold value for w in (3), we get:

$IR > b/2.$

From t = b/2 onwards, both agents will continue to play C: therefore, the degree of interpersonal relational intensity IR will grow more and more as time unfolds.

Hence, since partial altruism, far from being a fixed stock, evolves depending on the degree of relational intensity, we may characterize it as 'empathic altruism'. Empathy can be seen as "the ability to share the emotions of another person, and to understand that person's point of view" (Eysenck, 2000). Since altruism depends on the capacity to regard oneself as one individual among many, empathically-driven motivational types can be characterized as players who perceive what happens to others as if they were themselves involved in the situation. Several contributions have shown, mainly on the basis of psychological and biological reflections, that a significant correlation exists between perception of a person's distress and tendency to respond empathically through an overt helping act, that is to behave altruistically (e.g. Bateson, 1987; Hoffman, 1995). In their experimental analysis, Stahl and Haruvy (2006) also focus on the presence of empathy in the lab. They define empathy as follows: "since each participant is most likely a *recipient* rather than a dictator, the participant identifies with the recipient, invoking other-regarding preferences which otherwise might have been latent" (p. 28). Hence, their empathy hypothesis assumes that the weight assigned to other agents' monetary payoffs increases with the probability that you may be a passive recipient and they find that this empathy-altruism mechanism can explain the giving behavior in one of the experiments they run. In our work, empathy is triggered by sameness and, in turn, triggers altruism. Then, if the degree of altruism becomes sufficiently large, it may induce cooperative behaviors. This is the key mechanism at work whenever we assume that two altruists interact over time within a material PD framework. The empathic altruist differs from both the so called 'communitarian altruist' and the 'universal altruist': unlike the former, whose main feature is a selective attitude towards other people depending on their being embedded or not in his pre-formed social network, this motivational type selects others on the basis of *direct* social interaction, whereas he differs from the latter because of the key role played by his selective attitude towards others. Such behavior is likely to act as a good explanation of altruistic phenomena emerging within large groups, in which people do not know each other personally before social interaction occurs.

The same qualitative conclusion can be reached if we refer to a different functional form for U_A and U_B, assuming that both agents are not partial altruists, but are driven by 'Benthamite Altruism', in the sense that they care not only about their *own* payoff but also, in a traditional Benthamite fashion, about the *sum* of individual payoffs[10].

[10] Charness and Rabin (2002) focus on social welfare preferences characterized by agents who assign positive weight to aggregated surplus: other things being equal, the level of utility of an individual driven by such preferences increases if other agents are better off. They carried out 32 experiments and, by comparing several social preference approaches with the data, showed that social welfare preferences are better able to account for behavior in the experimental games.

The (linear) utility functions in this second case are:

$$U_A = (1-w_{AB})*\Pi_A + w_{AB}*(\Pi_A + \Pi_B) \qquad (4)$$

$$U_B = (1-w_{BA})*\Pi_B + w_{BA}*(\Pi_B + \Pi_A) \qquad (5)$$

It is straightforward to see that mutual cooperation (MC) arises here when t > b: now for the emergence of cooperation it takes twice the time required in the previous case. This difference can be easily accounted for by referring to players' utility functions; however, in both cases we notice that the larger is b, the greater is the level of t necessary for MC to endogenously emerge.

2.2. Motivationally heterogeneous players

In this section, in the light of the above analysis, we address the following question: what happens when we focus on *motivationally heterogeneous* populations? What about considering a more realistic world in which the previously illustrated co-evolutionary mechanism is supposed to be at work? Erlei (2006) interestingly remarks: "The overall impression is that the analytical combination of social preferences with heterogeneity in these preferences is a very productive way of understanding real behavior in laboratory experiments" (p. 21). Social preferences exhibit many patterns: reciprocally motivated players tend to display (seemingly irrational) nonstrategic punishment, often – though not exclusively – targeted towards defectors (see on this e.g. Fehr and Gaechter, 2000). As Erlei (2006) correctly observes, the growing behavioral economics literature on social preferences can be divided into three relevant substrands: theories of intention-based reciprocity (Rabin, 1993; Falk and Fischbacher, 2006), theories based on inequity aversion (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) and models centered around social welfare preferences (Andreoni and Miller, 2002; Charness and Rabin, 2002). In line with Erlei, our modelling strategy in this paper is very close to the latter two theories, in which social concerns are outcome-based and have to do with *distributional* issues, rather than with an evaluation of the opponent's intentions[11].

However, it appears sensitive to suppose that sociality may affect agents' utility either *positively* or *negatively*. This is confirmed by experiments on unselfish behaviors, showing that people are both willing to help others, caring about the others' material well-being or to be kind towards those who are kind (and such behaviors are often formalized through *pro-social* preferences such as altruism, inequity aversion or positive reciprocity) and unhappy if the opponent is better off, to be mean towards those who are mean or willing to explicitly punish others (and such behaviors are often formalized through competitive, *anti-social* preferences such as spitefulness, envy or negative reciprocity; see e.g. Fehr and Gaechter (2000) and Camerer (2003)). Levine (1998) correctly observes that while it appears plausible to suppose that people care not only about their own monetary payoffs, but also about their opponents' ones, it is not clear whether the coefficient on the other player's payoffs should be positive or negative: public good games seem to suggest that it is positive, but data about the well-known ultimatum game protocol indicate that it is negative (as positive offers are often rejected). As he does, we also assume that the coefficient differs between

[11] Levine (1998) adopts a somewhat intermediate point of view: while on the whole his approach is very close to distributional social preference theories, he elaborates a signalling game in which players' weights on opponents' monetary payoffs depend *both* on their own coefficient of altruism (or spite) and – in the spirit of psychological game theory (see Geanakoplos et al., 1989 and Rabin, 1993) – on *what they believe* their opponents' coefficients to be.

different agents in the population, with some players having positive coefficients and others having negative coefficients. Like Levine (1998), we further assume that each individual's coefficient is private information[12]. In our analysis, we assume that all the players' payoffs are linear in their own monetary income and their opponents', in the sense that all of them, albeit being motivationally heterogeneous, assign a positive relative weight (1-w) to their own material well-being and, under certain conditions, a positive relative weight (w) to either the opponent's material well-being or to the sum or the difference between their own and their opponent's material well-being. The latter component captures the *distributional* nature of their (either pro- or anti-) social preferences. Clearly, within this broader context, the parameter w will have to be interpreted as the degree of 'social influence', that is as the preference parameter characterizing the *content* of each type's distributional social preferences. This parameter captures the relative strength of distributive concerns compared to the purely individualistic ones and we assume that a certain degree of motivational heterogeneity exists with regard to the component that, within the utility function of a given type of players, multiplies such key 'parameter of sociality' w.

In the light of this, by using Matlab we have elaborated a simulation programme focusing on a heterogeneous population where six types of players driven by linear objective functions are considered. Motivational heterogeneity exists as our economy is populated by six player types, where three types of agents are driven by other-regarding preferences and three types are motivated by self-centered preferences. More specifically, while, as far as unselfish players are concerned, w multiplies either their opponent's well-being (Partial Altruism and, under certain conditions, Rawlsian Altruism) or the sum of their own and their opponent's well-being (Benthamite Altruism), other players are supposed to be driven by anti-social preferences: for them, w weighs the difference between their own and their opponent's well-being (Inequity Loving Egoism), at least under some conditions (Spiteful Egoism). Finally, some individuals are simply driven by classic self-regarding preferences (Pure Egoists), so that w = 0 for them.

Formally, our six player types' utility functions are as follows:

Partial Altruist: $U_i = (1-w_{ij})*\Pi_{ij} + wij*\Pi_{ji}$

Benthamite Altruist: $U_i = (1-w_{ij})* \Pi_{ij} + w_{ij}*(\Pi_{ji} + \Pi_{ij})$

Rawlsian Altruist: $U_i = (1-w_{ij})* \Pi_{ij} + w_{ij}* Ri$

with $R_i = \Pi_j$ if $\Pi_i \geq \Pi_j$

$= \Pi_i$ if $\Pi_i < \Pi_j$

Pure Egoist: $U_i = \Pi_{ij}$

Inequity Loving Egoist: $U_i = (1-w_{ij})* \Pi_{ij} + w_{ij}*(|\Pi_{ij} - \Pi_{ji}|)$

Spiteful Egoist: $U_i = (1-w_{ij})* \Pi_{ij} + w_{ij}* P_i$

with $P_i = \Pi_i$ if $\Pi_i > \Pi_j$

$= \Pi_i - \Pi_j$ if $\Pi_i \leq \Pi_j$

Beyond Pure Egoism, Partial Altruism[13] and Benthamite Altruism (whose basic features have been illustrated in the previous sections), we have decided to introduce three further

[12] As we clarify below, we suppose that a single player is informed about what his opponent has chosen to do in the previous interaction they had, but neither player knows what the opponent's type is.

[13] Both Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) assume that agents dislike inequality and are willing to sacrifice money in order to reduce it. As Erlei (2006) interestingly observes, both concepts of aversion to inequality

motivational structures in the population, so that it can be seen as more descriptively realistic: a Rawlsian Altruist assigns a positive weight to his opponent's payoff, but only insofar as the other player is either equal or worse off than him; an Inequity Loving Egoist is mainly interested in maximizing the 'degree of inequality' (formally expressed, here, by the difference between his own and his opponent's payoff); finally, the Spiteful Egoist's major feature is given by his being negatively affected by the difference between his own and his opponent's payoff, but only as far as he turns out to be worse off than his opponent. Hence, we also consider two types of players exhibiting *competitive* preferences, in the sense that both Inequity Loving and Spiteful Egoists positively weigh the opponents' negative payoffs[14].

Regarding Rawlsian Altruism, it is important to make clear that here we refer more to a recent and widely used interpretation of Rawls' maximin criterion rather than to Rawls' original idea itself, as the former is amenable to explicit game-theoretic treatment. In particular, with reference to such well-known mathematical formalization of Rawls' maximin, Alexander (1974) argues that such a concept, contrary to Rawls' claim, does not imply a complete departure from the standard utilitarian (Benthamite) framework; rather, the Benthamite and the Rawlsian criteria can be embedded into a common one-parameter family of welfare functions indexed by the relative weight assigned to individual utilities. This mathematical characterization of maximin allows one to consider a *continuum* of motivational structures placing different weights upon distributional concerns, of which maximin and Benthamite utilitarianism are but two focal points. As far as Inequality Loving Egoists are concerned, it is worth observing that while including such agents within the population under study may sound extravagant at first glance, their presence is far from implausible: such a motivational force seems to be commonly at work in all the societies where the very poor not only tolerate, but actively support the luxury of a small elite. We can interpret this as a 'preference for inequality' or as the implementation of a norm of elitarianism. Finally, Spiteful Egoism refers to those people who – far from being neutral towards other people's well-being – enjoy *negative psychological externalities* from the interaction with opponents getting an equal or larger material payoff[15]: here, these individuals are standard selfish agents whenever they are better off than their opponent, but are negatively affected by the difference between their opponent's and their own material payoff whenever they are worse off[16].

By including such motivational types within the overall framework, we create room for a potentially interesting comparison between the performance, at dynamic level, of (Inequality Loving and Spiteful) *anti-social* agents and (Rawlsian and Benthamite) *pro-social* individuals[17]. With respect to these 'extreme' figures, Pure Egoists are somehow in the middle, concerning

---

have been extremely successful in describing laboratory behavior when they assume *heterogeneous* actors. With reference to Fehr and Schmidt's specification of inequity aversion, it is easy to show that if we assume that the degree of aversion to unfavorable inequalities coincides with the degree of aversion to favorable inequalities, their objective function becomes very close to our formalization of partial altruism.

[14] It easy to notice that in such a society cooperation cannot become the only strategic choice made within the population due to the presence of agents that will systematically defect independently of the levels of both relational intensity and w

[15] For an interesting study on positionality, see Hirsch (1976).

[16] The main difference between these two player types driven by competitive preferences is that while Inequality Loving individuals are happier the larger is the degree of inequality between them and their opponent, regardless of their relative position compared to the opponent's one, the opposite occurs with regard to Spiteful Egoists: such players' utility *decreases* as their relative position worsens – compared to the opponent's one.

[17] Hence, it should be clear that we are not interested in Rawlsian (or Benthamite) social welfare theory *per se*; this is why we claim it is methodologically consistent to formally characterize Rawlsian players even though this is not what Rawls himself does. In other words, rather than taking an orthodox Rawlsian (or Benthamite, or Nietzschean, with regard to Inequity Loving players) perspective, we purposely construct some 'new', stylized motivational structures in order to see whether, in what cases, and under what conditions they can help to improve the implementable social outcome, once they are adopted within a complex population of heterogeneous agents.

exclusively (and 'coldly') about their own material well-being and totally disregarding the opponent's material payoffs.

3. Results

Before illustrating our main results, it is important to make clear that, with regard to the multi-player population under study, we introduce the following assumption, making the overall social environment even more 'hostile' to the emergence of cooperation. More specifically, unlike the analysis developed in Section 2 (regarding two-player populations only), we now suppose that, as far as pairwise matchings are concerned, whenever both individuals play D, each agent's level of relational intensity IR increases by one unit with probability h and *keeps constant* with probability 1-h. This stochastic element is likely to better capture the uncertain nature of the dynamics of human relations within complex societies, as it is reasonable to argue that it is often difficult to infer individuals' real intentions by simply observing their behavioral choices. In particular, it seems reasonable to think that, within a multi-player population, when a player's *specific* opponent defects, it is not easy for her to understand whether such behavior underlies 'bad' intentions towards her (i.e. an attempt of *exploitation*) or not[18]. It is important to recall that players are aware that they play a material PD whenever they meet another player, but also that the overall population is far larger than a two-player society in which the PD is 'the' overall game. We argue that this makes their interpretation of a mutual defection outcome different: if my opponent defects when I defect, this does not imply that he wants to be mean towards me, as we both know that we are part of a far larger society. In other words, I may see him as a 'free rider' towards society as a whole, rather than as a person who is trying to exploit me specifically. Hence, this similarity may well induce an increase in the level of the agent's 'parameter of sociality' w, that is his 'degree of social influence', regardless of the specific content of the player's distributional preferences[19]. Moreover, in order to not to favor, a priori, the emergence of cooperation, we rule out the possibility of mixed strategies: more specifically, we assume whenever agents are indifferent between the two strategies, they will defect.

The simulation has been carried out on a population of 60 agents. In the results we present, for expositional reasons, we forced the initial population to be composed by an equal number of all the types of the agents. Our main results hold also with a different distribution of population's agents. Parameters b, w and h have been kept constant, so that they did not affect the dynamics of the game. In particular, we have set b = 4.1[20] and h = 0.5. We also assume that whenever IR is negative w=0. The simulation is composed of 10.000 pair interactions ('row agent' versus 'column agent') and provides a matrix of strategies (called 'S'), a matrix of relational intensities ('IR') and a matrix for the levels of the social openness parameter w ('W').

The results illustrate what happens in the social scenario where different motivationally agents are involved. All the tables show the results emerging when the agent 'row' meets the agent

[18] Experimental economics has been increasingly showing the relevance of (one's opponent's) intention evaluation in motivating people's behavior (even when giving importance to intentions is materially costly). See on this Falk and Fischbacher, 2006.

[19] Also Tabellini (2008) in one of the versions of his model on the scope of cooperation assumes that individuals get a non-economic, psychological cost from defecting only if the opponent cooperates, but not if both players cheat. This is similar in spirit to our assumption that if two players cheat, then the level of relational intensity may increase.

[20] We have chosen not to use an integer number in order to avoid problems of indeterminacy with w, as such a parameter is related to IR through the expression w = IR/(IR+b).

'column'. They represent the spread of cooperation among the 6 types of agents (Figure 1), the evolution of IR (Figure 2) and the evolution of W (Figure 3) over time, respectively.

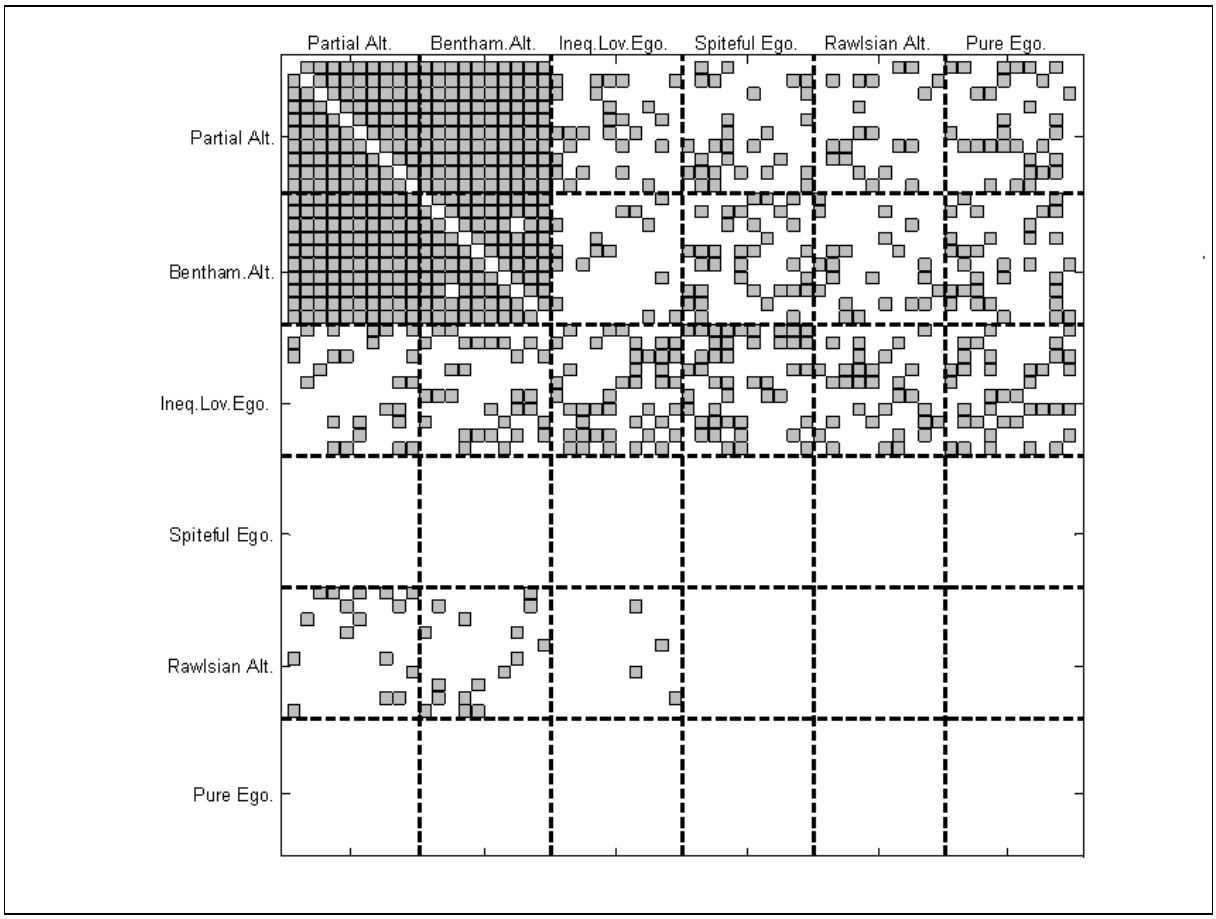

Figure 1. Final frequency of cooperation

Result 1. *Mutual cooperation stably emerges among Partial Altruists, Benthamite Altruists and between them*

As far as players driven by pro-social preferences are concerned, we see that when Partial Altruists interact with one another, cooperation becomes the unique strategy (fig 1) and the values of IR and W are very large (fig 2 and 3). The same is true with reference to Benthamite Altruists: in this case cooperation is widespread. Stable cooperation also emerges between these two types of agents. Even if the period in which they change their strategy (from defection to cooperation) differs due to the differences in the utility functions, in the long run the Partial Altruist has to engage in a longer period of mutual defection in order to allow the social openness of the Benthamite altruist to be sufficiently high to bring about a change in the latter's strategies towards Partial Altruists. It is also interesting to notice that those two types of agent systematically adopt a cooperation strategy towards all the other types of agents. These 'cooperation attempts' do not succeed in building a stable relation based on cooperation because they never match another cooperation strategies from the opponents. Therefore, IR and W decrease due to the difference in behavior, and in the following period Partial Altruists and Benthamite Altruists will defect again.

This difference in behavior is reflected in the fact that the relational intensities of these two types with respect to the other four types of agents are almost zero all along the game.

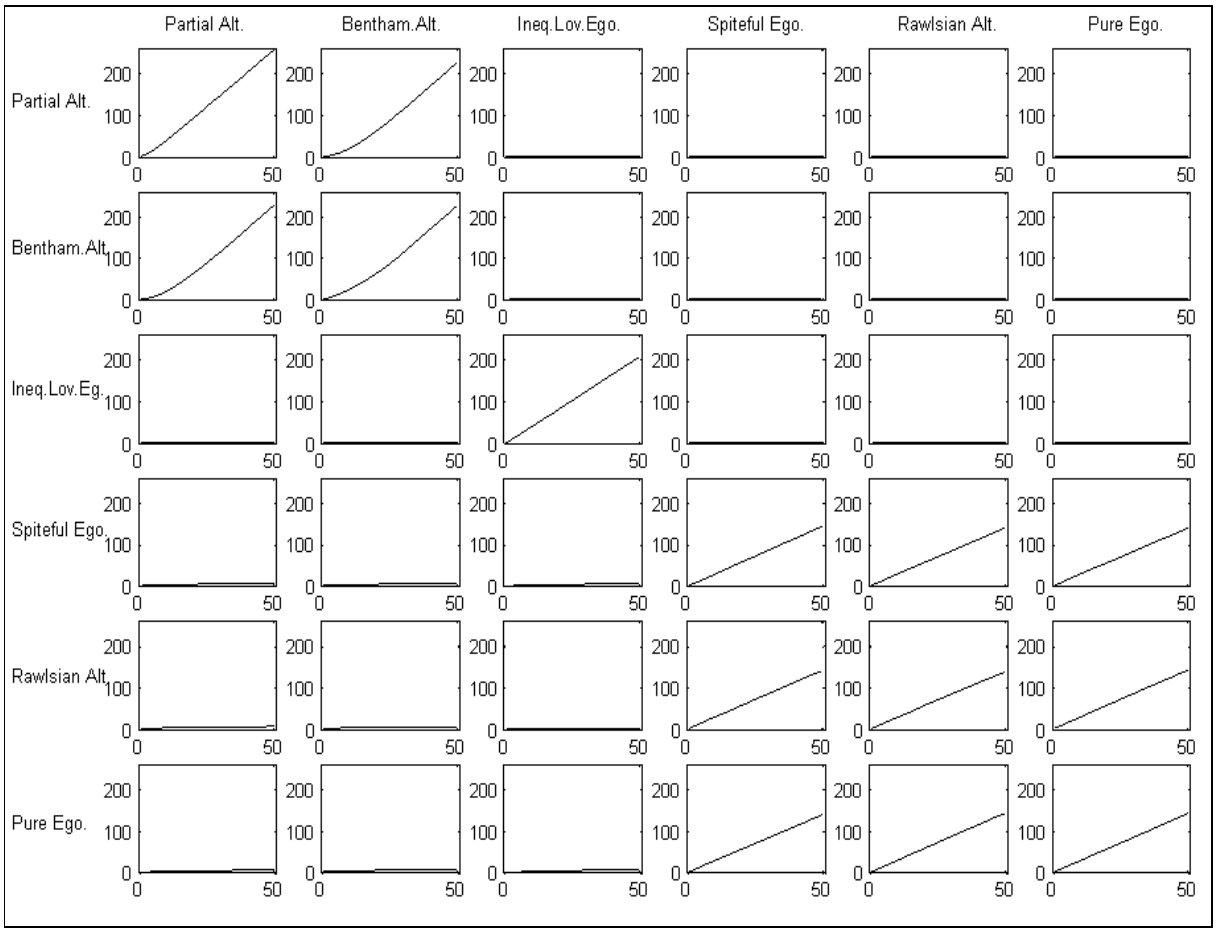

Figure 2. Evolution of the average of the Relational Intensity parameters (IR)

Result 2. *A cyclical behavior characterizes Inequity Loving Egoists*

The case of Inequity Loving Egoists is very interesting: among them they cyclically play mutual defection, which is followed by a period of mutual cooperation. Due to the symmetry of the strategy among them, the values of IR and W increase (See figures 2 and 3). This happens because (for sufficiently high values of IR and W) they reach a point in which changing strategy gives them a higher utility than the previous strategy, due to the fact that, in addiction to their own monetary payoff, they also maximize the absolute value of differences in payoffs. Inequity Loving Egoists also adopt a cyclical strategy against all the other types of agents: more specifically, they always choose the opposite strategy, compared to the one previously played by their opponent. Let us note (Figure 2) that in this case their IR with respect to the other types of opponents is always close to zero, right for the reason that they never coincide in the strategies.

With regard to Spiteful Egoists, independently of the values of relational intensities, social openness and the type of opponent, they never cooperate as a response to cooperation. Also due to the fact that they positively weight social distance only when the opponent are worse off (this

differs from Inequity Loving Egoists that appreciate social differences independently of who obtains the worse result) they never cooperate against defection. Their IR is basically zero with respect to Partial Altruists, Benthamite Altruists and Inequity Loving Egoists due to the constant differences in behavior. On the other side, IR and W constantly increase, on the basis of mutual defection, when we consider the interaction among Spiteful Egoists and the one between Spiteful Egoist with Rawlsian Altruists and Pure Egoists. In fact it is worthy to notice that those values are nearly half of the ones characterizing Partial and Bentimite Altruists, which increase their relational intensity towards mutual cooperation (Figure 2).

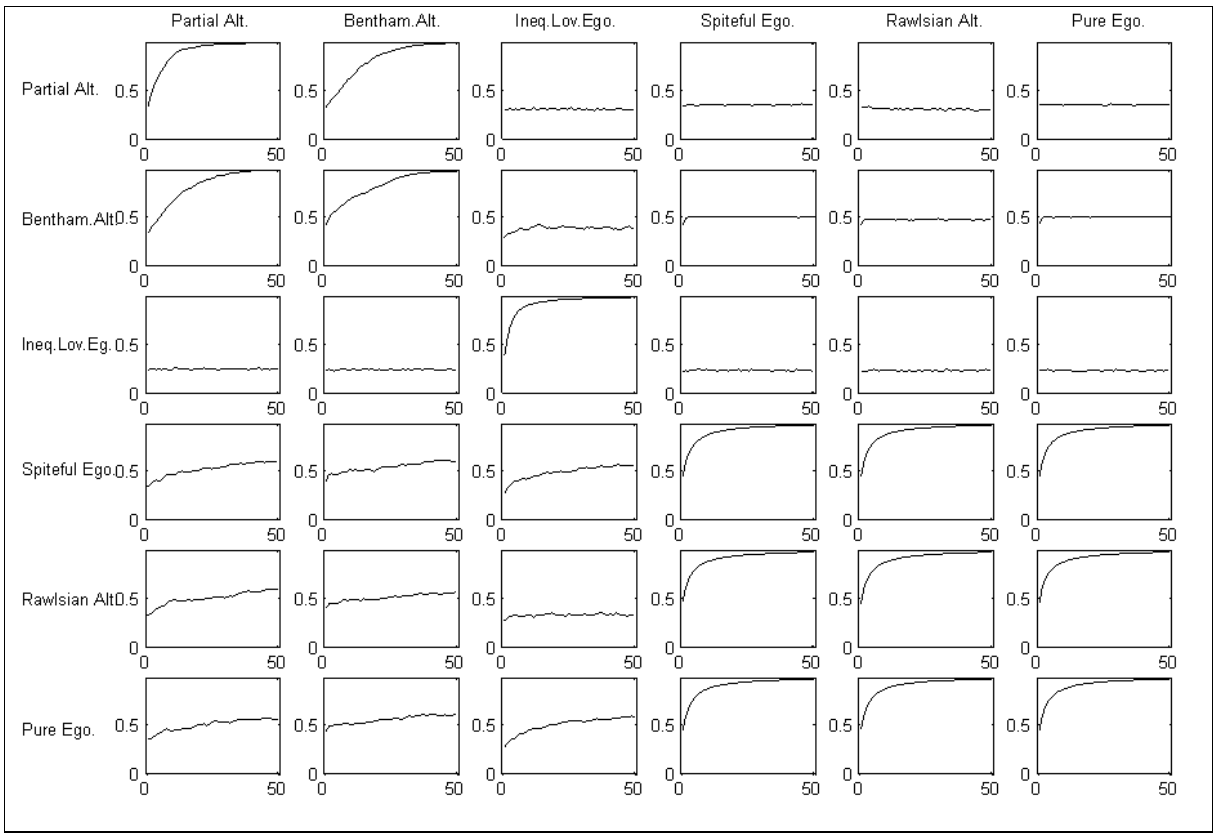

Figure 3. Evolution of the average of social openness parameters (W)

Pure Egoists, by definition, always defect and their only positive values of IR increase based on mutual defection.

Rawlsian Altruists, among themselves, tend to behave as Spiteful Egoists and Pure Egoists do (and the values of W and IR are similar). The reason is as follows: they are never the first to cooperate, since for them it is rational do so, only as a response to previous cooperation, therefore they cannot succeed in cooperate among them. Notwithstanding this, IR and W increase based on mutual defection. On the other hand, due to the fact that they weigh opponents' payoff when those are less then their own payoff, they cooperate with Partial and Benthamite Altruists and with Inequity Loving Egoists, after the period in which those players had cooperated first. But, since the periods of cooperation never coincide, cooperation fails to emerge.

3.3. Evolutionary dynamics and IR-based interactions: an indirect evolutionary approach

Gueth and Kliemt (1998) assert that it is not plausible to take individual preferences as exogenous to social interaction; by contrast, preference formation should be integrated into economic models. In the same vein, as Tabellini (2008) observes, while social sciences like sociology have often discussed the endogenous evolution of values and preferences, economics has normally taken individual preferences as given, without focusing on the endogenous evolution of preferences, norms and values. As he remarks: "In many social situations individuals behave contrary to their immediate material self-interest, not because of an intertemporal calculus of benefits and costs, but because they have internalized a norm of good conduct" (p. 2). He also wonders: "Why do specific values persist in some environments and not in others?" (p. 2). The evolutionary game-theoretic approach provides a natural environment for the analysis of 'motivational ecologies' where heterogeneous players interact over time. The idea of a 'motivational ecology' stresses the fact that in this context alternative motivational systems act as if they were "struggling for survival", i.e. compete in order to be adopted by the largest possible number of players. Within such an environment, it is possible to set up dynamic models of social learning describing the evolution of behaviors generated by social interaction processes and explain how a specific subset of the original choice set is eventually 'selected' in a self-enforcing way by the social dynamics. The theoretical framework just described is in principle compatible with several different learning mechanisms, as it seems to emerge from the important literature on the learning 'microfoundations' of evolutionary dynamics. It is worth observing that while the range of possible social conventions resulting from the process is determined by the socially established choice set (i.e. by the set of motivational structures that players consider as viable alternatives), the convention that actually emerges depends entirely on the dynamic interaction of individual choices. This compromise may be the basis of a reasonable medium-run modelling approach to the interplay of the sociologically and economically oriented components of human action: people act as imperfect optimizers, but only within the choice context that is provided by the social and cultural environment they are embedded in.

In this light, the new evolutionary assumption that we add to the model and introduce in this section is that the agents that obtain the best results are able to replicate themselves more quickly than the others and, as a result, can increase their presence within the population. In particular, we assume that agents can periodically *observe* the *average* monetary payoff of the entire population and compare this value with their *own* average expected monetary payoff. If an agent's expected result is worse than this value, she will change type with a probability proportional to the average payoffs obtained by all the types. Therefore, the higher are the expected payoffs of one type, the higher is the probability that the other will mimic his behavior. This mechanism allows for the possibility that an agent could remain of the same type since all the average expected payoffs are taken into account. It is important to point out that an agent does *not* compare *utilities*, but simply monetary payoffs: hence, our approach is fully in line with the well-known 'Indirect Evolutionary Approach' (IEA) pioneered by Gueth and Yaari (1992). In the IEA, preferences are treated as endogenous to an evolutionary process in which "objective evolutionary success depends on the choices made, which in turn depend on subjective preferences. Success feeds back on subjective preferences, and so on" (Gueth and Kliemt, 1998; p. 377). In this evolutionary version of our model, we adopt a *co-evolutionary* approach which may be seen as a variant of the IEA[21], as we treat preference formation as a result of a social interaction process in which subjective preferences drive choices and both (i) *behavioral choices* (due to the key mechanism linking behavioral choices, relational intensity and degree of social influence) and (ii) *monetary payoffs* feed back on subjective preferences over time. Outcome-based preferences here are endogenous as "human choice behavior

[21] Also Janssen (2006) adopts the IEA.

is not exclusively determined by expectations of future consequences of choice alternatives. (…) We are pushed by innate as well as acquired dispositions and at the same time are pulled by future directed expectations and desires, which result in corresponding intentions. (…) an indirect evolutionary approach along with preference formation can simultaneously incorporate both 'dispositional push' and 'expectational pull'" (Gueth and Kliemt, 1998; p. 378 and 380). In this paper we stick to a similar approach: 'phenotypes' interact in pairs and play a material PD and, when doing so, they act 'rationally' on the basis of their subjective preferences. In other words, in pairwise matchings their choices are driven by the 'expectational pull': rationality is *forward-looking* and drives choices, on the basis of a lucid calculus of *expected* material consequences. At this micro level (single rounds of play involving two agents at a time), rationality plays a key role: individuals are consequentialistically oriented, preferences can be reasonably taken as given and purely preference-based behavior occurs[22]. However, as time unfolds the above described co-evolutionary process goes on and also the 'dispositional push' needs to be taken into account, as preferences themselves are supposed to *gradually evolve*, due to both (i) the mechanism linking behavior, relational intensity and degree of social influence and (ii) material success. Hence, on the whole our co-evolutionary model, like the IEA, incorporates both 'dispositional push' and 'expectational pull': while the latter drives choices in each matching, the former makes the 'engine' of choices (that is, subjective preferences) endogenous to social interaction. Regarding the 'evolutionary engine' being at work within our framework, it is worth trying to understand whether it is true or not that, since material consequences are supposed to play a key role, selective incentives actually play against individuals who pay attention to other people's well-being and that, as a consequence, these motivational types should sooner or later die out in any specific socio-economic environment. Unlike the prediction underlying such 'naïve evolutionary view', we show that selective incentives need not always favor self-interested individuals at the expenses of nonselfish ones.

We explore this by also supposing that for each player the probability of being matched with another player positively depends on the degree of relational intensity characterizing their interactions. This assumption appears plausible and gives us the chance to observe the presence of *clusters* within the overall population. Specifically, we assume that the probability with which the agents are matched is proportional to the degree of relational intensity IR that an agent experiences with her opponents, according to the following formula:

$$(12)\quad p_{ij} = IR_{ij} / \sum_j IR_{ij}$$

This seems a reasonable assumption that captures the idea that agents prefer to interact with opponents that showed similarity in behavior, which is the key determinant of increases in relational intensity. We can think of this relationship between matching probabilities and IR as a mechanism due to which behavioral differences are 'punished'. However, it is important to emphasize that such mechanism is *not* ex ante biased in favor of cooperation and against defection, as it implies that, in the presence of mutual defection (and not only of mutual cooperation), IR increases and so does the matching probability. In a similar vein, Eshel, Sansone and Shaked (1999) set up a model in which agents exclusively interact with a small subset of the overall population (i.e. with their neighbors only) and in this framework they are able to show that if a strategy is 'unbeatable' (that is, robust against the possible invasion of a finite group of identical mutants), then such a strategy is unique and is given by altruism (namely, agents turn out to behave as if they were related to their neighbors and take into account their welfare as well as their own payoffs). In particular, they find that the

[22] Hence, as Gueth and Kliemt correctly remark, within each round conventional game theory can be fruitfully applied in order to predict the results of social interaction.

degree of altruism depends on the ratio between the radii of the interacting and the learning neighborhoods.

3.2. Major results: nice guys finish first

In this subsection, we summarize our major results, with regard to the evolution of player types and the dynamics of material payoffs (Result 3), the creation of significant interpersonal relationships (Result 4) and the frequency of cooperation (Result 5) within the final population emerging from the simulatory analysis carried out along the lines clarified above. Once again we forced an equal number of the six types of the initial population. Results hold also in the random population cases.

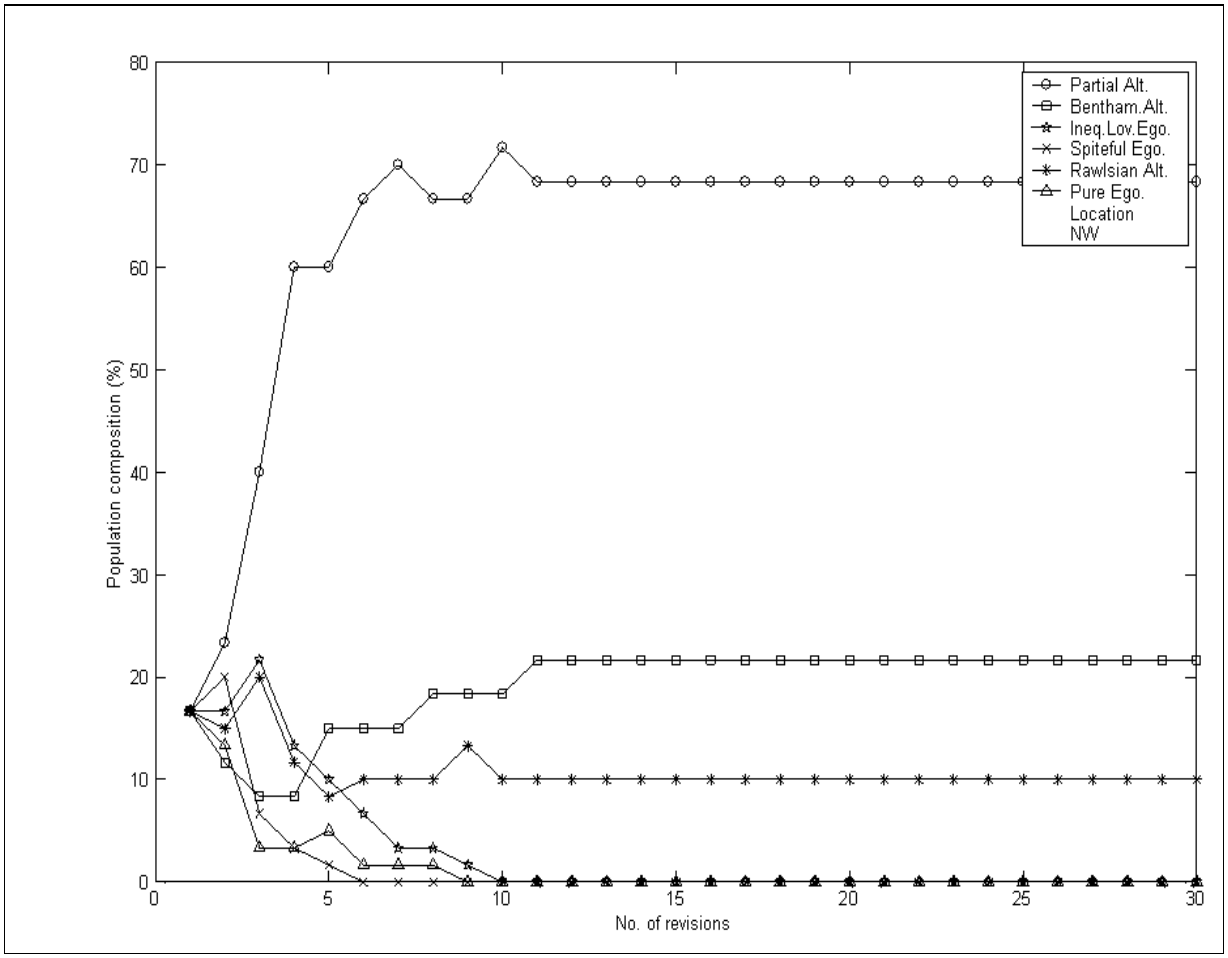

Figure 4. The evolution of player types

Result 3. *Egoists disappear (Inequity Loving Egoists disappear first, followed by Pure Egoists and, subsequently, by Spiteful Egoists) first*

With very strong regularity, a sequence occurs in which the three types of egoists will disappear within the population: the Inequity Loving Egoists are always the first to change their type, because of their behavior: as a result of their decision to always make the opposite strategic choice compared to that of their opponents at the beginning of the game, when everyone defects,

they will get an expected payoff which is always less then one. This does not happen to altruists, because they can gain from cooperating among themselves. The second type of selfish players who disappear is give by Pure Egoists. Even if they gain from exploiting the behavior of the altruists, the assumption concerning the non-random nature of matchings makes the interactions with them relatively rare. Hence, they end up always defecting against the other types of egoists. Later this happens also to Spiteful Egoists, but they can survive longer because they tend to play cooperatively more often than the other egoists do.

Further, if we look more closely to the final composition of the population, we notice that, whereas Benthamite and Rawlsian Altruists are always present when cooperation spreads over within the entire population, Partial Altruists are not always present. We see that roughly in one case out of three Benthamite and Rawlsian Altruists are the only player types that remain in the game. This is due to the fact that Partial Altruists are the ones who will cooperate first with all the agents and, at the beginning of the game, they will never gain by making this behavioral choice. Insofar as there is not a sufficiently large number of Partial Altruists around (among which cooperation could flourish), the results that such players obtain are clearly worse than the ones obtained by all of their opponents opting for defection against them. However, regardless of the presence of two or three types of altruistic individuals in the final population, cooperation turns out to be the only strategic choice that survives, leading to an equal average payoff (corresponding to the (C,C) equilibrium) among altruists (see Figure 5).

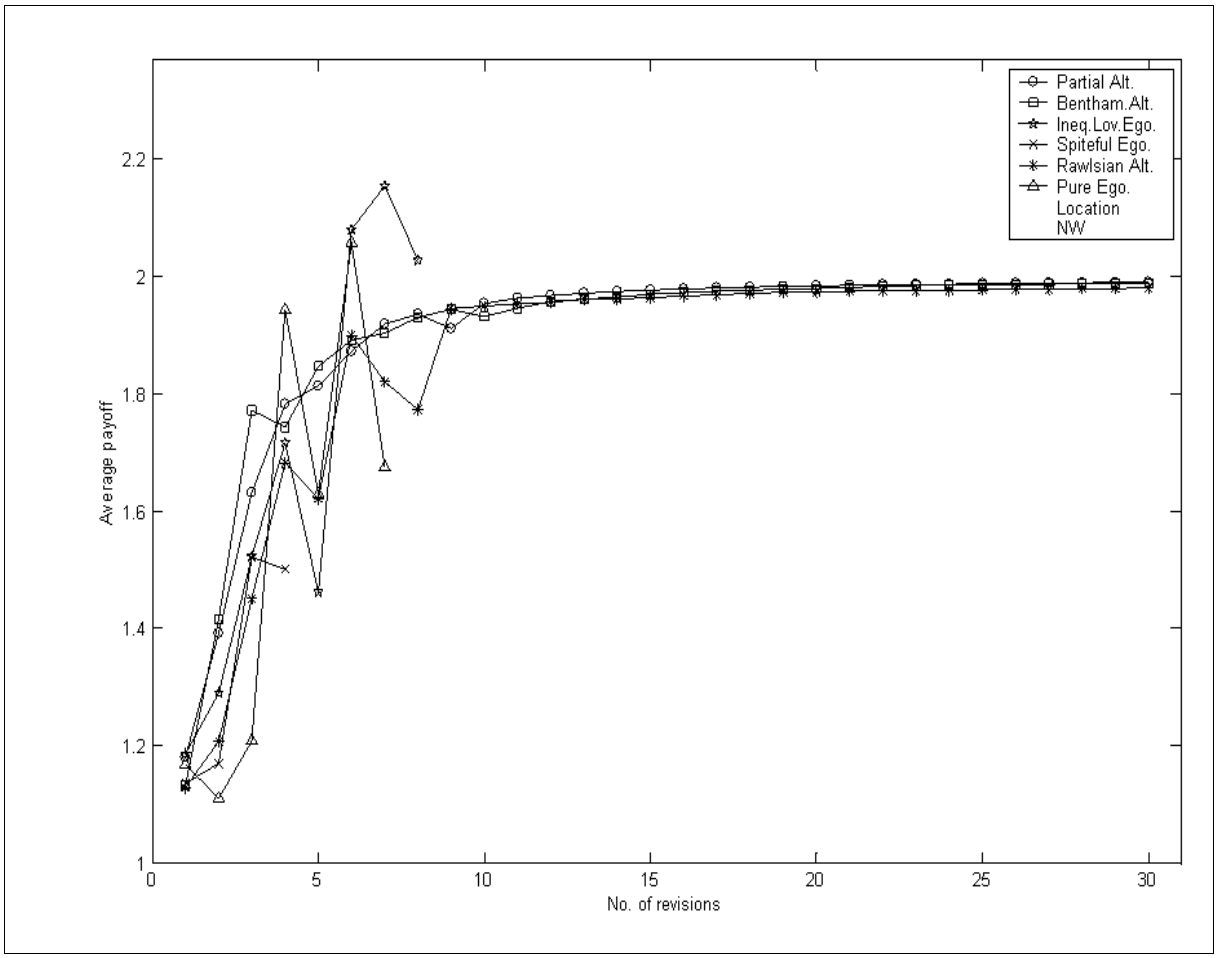

Figure 5. The dynamics of average material payoffs (Egoists disappear)

By analyzing the dynamics of interpersonal relationships, we find extremely robust results. Altruists generate very strong relationships among themselves through cooperation. In fact, as time unfolds, the only positive values of IR (Figure 6) are the ones among the three types of Altruists, with the only exception of the Relational Intensity among Rawlsian Altruists, who, as we explained in the previous paragraph, are not able to cooperate among them. Moreover, it is important to notice that, due to the fact that the matching probabilities depend on the values of IR, Rawlsian Altruists have almost zero probability of interacting among them, due to the fact that their Relational Intensity remains very low. In the long run they establish stable mutual cooperation only with the other two types of Altruists (Figure 7). This can be summarized as follows:

Result 4. *As time unfolds, altruists establish increasingly strong relationships among themselves, based on mutual cooperation*

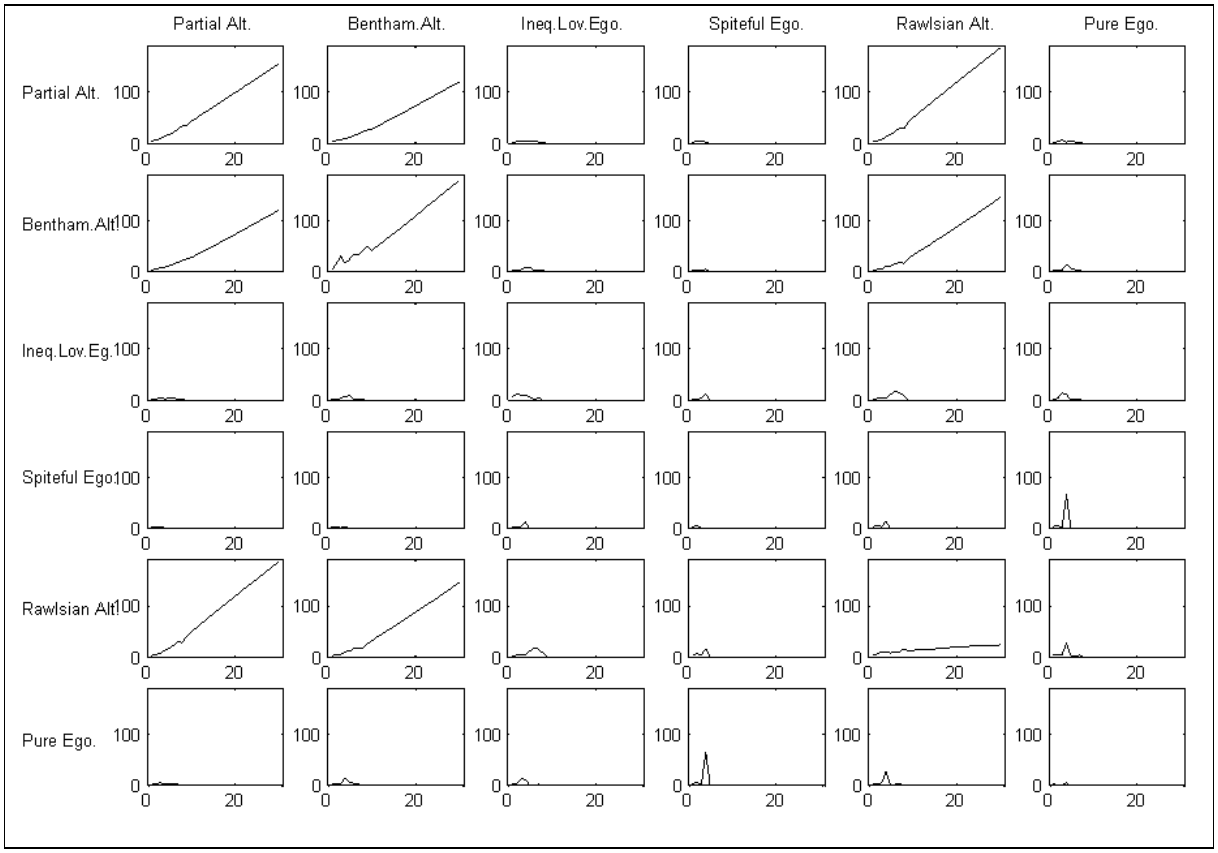

Figure 6. Dynamics of average IR among types

This results in roughly double values of relational intensities compared to the case of egoists who strengthen their relationship on the basis of defection. If we look at Figure 3, these dynamics are very clear. Here we reported the average levels of relational intensity (IR) among groups, where the types are ordered in the same way as the results concerning the homogeneous types. For example, the first row of graphs summarizes the evolution of the average level of IR of Partial Altruists with respect to themselves, Benthamite Altruists, Inequity Loving Egoists, Spiteful Egoists, Rawlsian

Altruists and Pure Egoists. Hence, it is straightforward to notice that the levels of IR become increasingly large in the first, second and fifth graph in rows 1, 2 and 5, that is with regard to the three types of altruists present within the overall heterogeneous population.

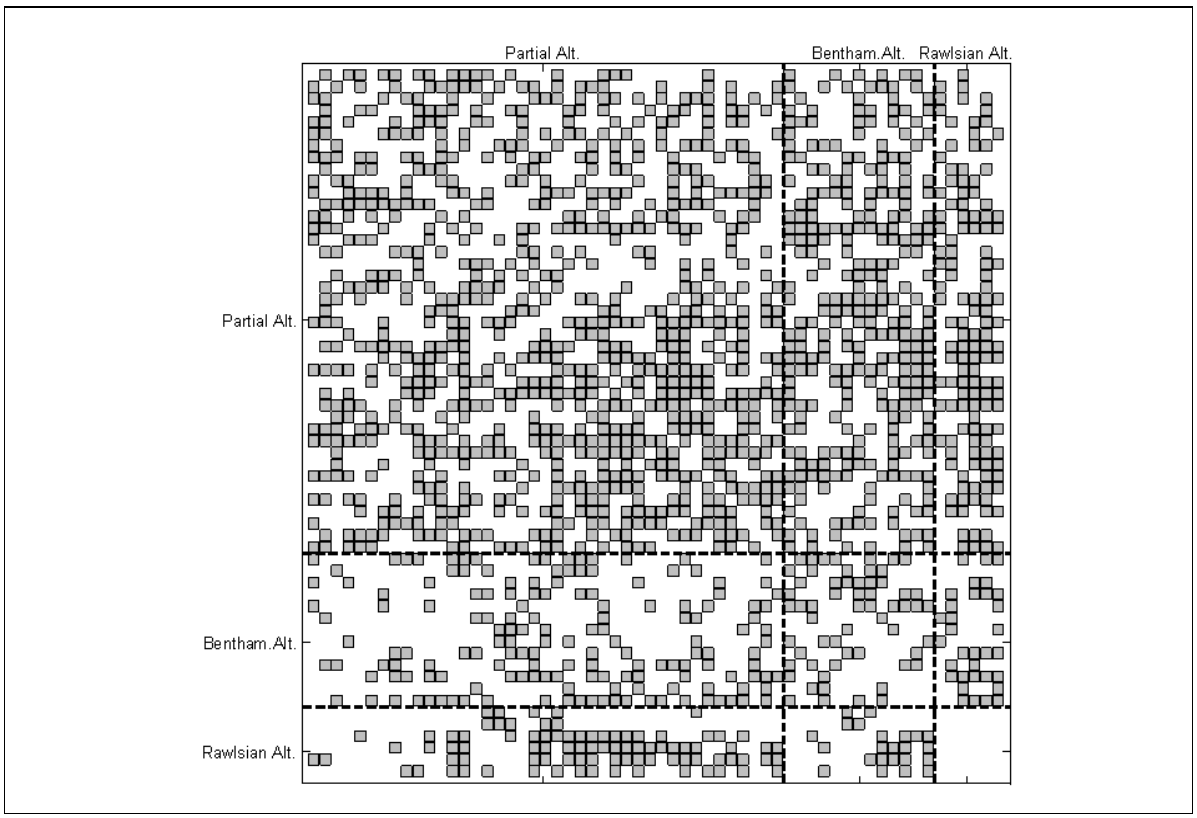

Figure 7. The frequency of cooperation

Finally, in Figure 7 we report the frequency of cooperation among the agents that survive. Here, a dot represents the cooperative strategic choice of the agent 'row' with respect to agents 'column'. As we pointed out before, cooperation is the only 'active' strategic choice within the final population (as it is evident from observing the average payoff, which is equal to two). The empty spaces (which strictly speaking indicate defection), here can be interpreted as the agent's relationships that are 'inactive', in the sense that defection is still the best strategic choice against agents with whom the probability to be matched is close to zero. The results concerning the evolution of player types and the dynamics of material payoffs captured by figures 5 and 7 can be summarized as follows:

Result 5. *Within the final population, cooperation is the unique strategic choice played within the social network*

4. Concluding remarks

In Sections 2, 3.1. and 3.2. we have illustrated a model where different types of individuals are connected in a circular way: in the short run, in pairwise matchings, maximization of utility functions drives the behavior of agents that has, not only (direct) consequences in terms of material payoffs, but also (indirect) consequences – through players' behavioral choices – on the degree of relational intensity IR. IR affects the motivational dimension of the subjects through the parameter of sociality. Hence, the model underlies a co-evolutionary approach based on the interplays among motivational, behavioral and relational level, with regard to both homogeneous and heterogeneous populations. In Section 3.3. we retain this basic framework but we add two key assumptions, with regard to evolutionary dynamics: by adopting a variant of the indirect evolutionary approach, we suppose that material payoffs play a critical role in driving the long-run evolution of types. Moreover, we assume that here IR affects not only players' utility functions (via the degree of social influence) but also the probability with which the agents are matched.

The simulations we carried out within such more complex evolutionary environment led to strongly consistent results with regard to both the evolution of the types and the creation of significant relations among agents. In the long run, cooperation turns out to be the strategic choice that obtains the best results, in terms of material payoffs: as a consequence, the large majority of the populations we analyzed is composed of *altruists only* (this regards more than 95% of the cases). Hence, nice guys, far from finishing last, prevail and dominate within the final population, where they establish significant interpersonal relationships among themselves. More specifically, unlike the case previously analyzed and discussed, Rawlsian altruists are now also able to adopt their preferred behavior (cooperation) in a stable way and they succeed in generating extremely strong relationships with the other two types of altruists (see Figure 1 above). It is important to emphasize that there we reported the results of a single simulation capturing the majority of our general results[23]. In the remaining cases (5%), the final population consists, with equal probability, of the three types of egoists and the only strategic choice that emerges in the population is defection. However, this result is mainly due to the fact that, within the initial population, very few altruists (of all the kinds) were present and, at the beginning of the game, they obtained very bad results due to the low number of agents with whom they could cooperate. Thus, their altruistic behavior has been fully exploited by their opponents.

[23] The results of the other simulations not reported here are available upon request.

References

Alexander, S., 1974. Social evaluation. Quarterly Journal of Economics 88. 597-624.

Andreoni, J., Miller, J., 2002. Giving according to GARP: an experimental test of the consistency of preferences for altruism. Econometrica 70, 737-753.

Axelrod, R., 1984. The emergence of cooperation among egoists. The American Political Science Review 75.

Batson, C.D., 1987. Prosocial Motivation: Is It Ever Truly Altruistic? In L. Berkowitz (Ed.) Advances in experimental social psychology, 20, Academic Press, New York.

Benchekroun, H., Van Long, N., 2007. The build-up of cooperative behavior among non-cooperative selfish agents. Journal of Economic Behavior and Organization. Forthcoming.

Bolton, G., Ockenfels, A., 2000. A theory of equity, reciprocity and competition. American Economic Review 100, 166-193.

Camerer, C., 2003. Behavioral Game Theory. Princeton University Press, Princeton.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117, 817-869.

Edgeworth, F.Y., 1881. Mathematical Psychics, Kegan Paul and Co., London.

Erlei, M., 2006. Heterogeneous social preferences. Journal of Economic Behavior and Organization, forthcoming.

Eysenck, M.W., 2000. Psychology. A Student's Handbook, Psychology Press, Hove, East Sussex.

Eshel, I., Sansone, E., Shaked, A., 1999. The emergence of kinship behavior in structured populations of unrelated individuals. International Journal of Game Theory 28, 447-463.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293-315.

Fehr, E., Fischbacher, U., Kosfeld, M., 2005. Neuroeconomic foundations of trust and social preferences. American Economic Review 95, 2, 346-351.

Fehr E., Schmidt K., 1999. A theory of fairness, competition and co-operation. Quarterly Journal of Economics 114, 817-868.

Fehr, E., Gaechter, S., 2000. Cooperation and punishment. American Economic Review 90, 4, 980-94.

Gintis, H., Smith, E., Bowles, S., 2001. Cooperation and costly signalling, Journal of Theoretical Biology 213, 103-119.

Güth, W., Yaari, M., 1992. Explaining Reciprocal Behavior in Simple Strategic Games: an Evolutionary Approach, in Witt U. (ed.), Explaining Process and Change: Approaches to Evolutionary Economics, Ann Arbor, University of Michigan Press, 23-34.

Hamilton, W.D., 1964. The genetical evolution of social behavior. Journal of Theoretical Biology 7, 1-16.

Hoffman M.L., 1995, Personality Processes and Individual Differences - Is Altruism Part of Human Nature?, in Zamagni S.(a cura di), 1995.

Janssen M.A.. 2006, Evolution of cooperation in a one-shot prisoner's dilemma based on recognition of trustworthy and untrustworthy agents. Journal of Economic Behavior and Organization. Forthcoming.

Li, J., 2006. The power of conventions: A theory of social preferences. Journal of Economic Behavior and Organization. Forthcoming.

Mueller D., 1986, Rational Egoism versus Adaptive Egoism as Fundamental Postulate for a Descriptive Theory of Human Behavior, Public Choice, 51, 3.

Nowak, M.A., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. Nature 393, 573-77.

Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281-1302.

Rotemberg, J.J., 2007. Minimally acceptable altruism and the ultimatum game. Journal of Economic Behavior and Organization. Forthcoming.

Rubinstein, A., 1986. Finite automata play the repeated Prisoner's Dilemma. Journal of Economic Theory 39, 83-96.

Singer, T., Fehr, E., 2005. The neuroeconomics of mind reading and empathy. American Economic Review 95, 2, 340-345.

Smith A., 1759, The Theory of Moral Sentiments, (ed.) Stewart D., Bell and Sons, 1892, New York.

Stahl, D.O., Haruvy, E., 2006. Other-regarding preferences: Egalitarian warm glow, empathy, and group size. Journal of Economic Behavior and Organization 61, 20-41.

Tabellini, G., 2008. The scope of cooperation: values and incentives. Quarterly Journal of Economics. Forthcoming.

Trivers, R.L., 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46, 1, 35-57.

Weibull, J., 1995. Evolutionary Game Theory. MIT Press, Cambridge and London.