Zamagni, Stefano; Sacco, Pier Luigi

Working Paper

An Evolutionary Dynamic Approach to Altruism

Quaderni - Working Paper DSE, No. 165

Provided in Cooperation with:
University of Bologna, Department of Economics

Suggested Citation: Zamagni, Stefano; Sacco, Pier Luigi (1993) : An Evolutionary Dynamic Approach to Altruism, Quaderni - Working Paper DSE, No. 165, Alma Mater Studiorum - Università di Bologna, Dipartimento di Scienze Economiche (DSE), Bologna, https://doi.org/10.6092/unibo/amsacta/5192

This Version is available at:
https://hdl.handle.net/10419/159008

An Evolutionary Dynamic Approach to Altruism*

Pier Luigi Sacco
Department of Economics
University of Florence

and

Stefano Zamagni
Department of Economics
University of Bologna

June 1993

JEL classification numbers: C72, C73, D74.

Abstract

After reviewing recent and less recent literature on altruism and selfishness, the paper discusses in detail the issue of the relevance of altuistic motivations, as opposed to self- interest, in a well grounded theory of rational choice. The main point is that a proper dynamical analysis of the allocative consequences of altruism must focus not only on the actual level of altruism that characterizes the economy, but also on the types, keeping in mind that the time evolution of the stock of altruism crucially depends on the distribution of the various types across the population. It is shown that the model here presented can be used to interpret altruistically - driven phenomena of transation between social conventions.

---

* Financial support through MURST 40% funds is gratefully acknowledged.

"A man who is a mathematician and nothing but a mathematician may live a stunted life, but he does not do any harm. An economist who is nothing but an economist is a danger to his neighbours. Economics is not a thing in itself; it is a study of one aspect of the life of man in society... Modern economics is subject to a real danger of Machiavellism — the treatment of social problems as matters of technique, not as facets of the general search for the Good Life" (Hicks (1941)).

1. Introduction.

A social convention is, according to the classical definition of Lewis (1969), a state of things which is customary, expected and self–enforcing. In game theoretic terms (as pointed out e.g. by Ullman–Margalit (1977)), these requirements can be translated into the familiar Nash equilibrium conditions. Very recently, Sugden (1989) has pointed out that historically relevant social conventions need not only be self–enforcing vis–a–vis themselves but also against the possible 'invasion' of 'rival' conventions, i.e. against the possibility that a relatively small number of players act as 'deviants'. This additional feature amounts to a refinement of Nash equilibrium known as evolutionary stability (see e.g. van Damme (1987)). This characterization of social conventions clearly stresses their stability, i.e. the fact that, once the social convention has been established, its destruction requires the successful coordination of a relatively large number of players. Unfortunately, as it is well known, it is often the case that social conventions do not correspond to socially desirable outcomes; this can be due to the fact that:

a) there exist many possible social conventions, but for some reason the 'wrong' one has become salient. One such example is the so called coordination failure: workers coordinate on an inefficiently low level of activity despite the fact that a *uniformly* higher level of activity would make all them better off (see e.g. Cooper and John (1988));
b) the only feasible social convention is Pareto inferior to some other, non implementable outcome. The most famous example in this respect is the prisoner's dilemma game; another well known situation

is the so called tragedy of the commons (see e.g. Hardin (1968) but also Bromley (1991)).

Even when socially preferable conventions other than the ruling one exist, the transition from the latter to the former is problematic for the reasons hinted at above. To explain the transition between social conventions one has therefore to invoke some additional feature that forces the economy away from the original situation. Boyer and Orlean (1992) provide an example of transition between conventions based on a certain pattern of localization of strategic interaction between players. More examples based on different factors may be devised; in this paper, we are particularly interested in one of them, namely altruistic behaviour.

To understand how the emergence of altruistic behaviour may cause the transition between different social conventions or even the emergence of a desirable convention when there is none, consider, as a simple example, the following specification of the standard prisoner's dilemma game:

	C	D
C	(ς, ς)	$(\varsigma - \psi, \varsigma + \lambda)$
D	$(\varsigma + \lambda, \varsigma - \psi)$	$(0, 0)$

where $\varsigma, \psi, \lambda > 0$, $\psi > \varsigma$.

As it can be easily checked, (D, D) is the only Nash equilibrium here, and hence the only enforceable social convention. If on the other hand at least one player (say, the row player) is altruistic in the sense that her target payoff function is a weighted average of her own and the opponent's payoff, it turns out that cooperation is a dominant strategy for her provided that w, i.e. the weight assigned to the opponent's payoff, is large enough. More precisely, if $\varsigma + \lambda > \psi$, it is required that $w > \lambda/(\lambda + \psi)$, whereas if $\varsigma + \lambda < \psi$, it is required that $w > (\psi - \varsigma)/(\lambda + \psi)$.

If on the other hand *both* players are altruistic, mutual cooperation may become the only enforceable social convention. It might be objected that the emergence of mutual cooperation as a social convention requires a relatively high degree of altruism. It can indeed be shown that, when

$\varsigma + \lambda < \psi$, even for lower degrees of (row player's) altruism (namely, for $\lambda/(\lambda+\psi) < w < (\psi - \varsigma)/(\lambda + \psi)$) the original prisoner's dilemma can be transformed into a so–called assurance game in which free riding is not the most attractive option for the altruistic player (who now ranks outcomes as follows: $(C,C) \succ (D,C) \succ (D,D) \succ (C,D)$; see e.g. Sen (1967)). The payoff matrix then becomes (with the row player as the altruist):

	C	D
C	(ς, ς)	$(\varsigma - \psi + (\lambda + \psi)w, \varsigma + \lambda)$
D	$(\varsigma + \lambda - (\lambda + \psi)w, \varsigma - \psi)$	$(0,0)$

In this case, the altruistic player has no longer a dominant strategy; it is often suggested (see e.g. Elster (1984)) that defection is the rational strategy for her to play. However, as argued by Collard (1978), the row's player assurance about the opponent's choice of (cooperative) strategy need not be complete in order for her to choose to behave cooperatively. The altruistic player will indeed cooperate as long as her subjective estimate p of the probability that the opponent will cooperate is large enough[1]; it is easily checked that, the larger the player's altruism, the less assurance she requires to justify her own cooperation[2]. If on the other hand *both* players are sufficiently altruistic (i.e., $w > \lambda/(\lambda + \psi)$), the cooperative outcome is again self–enforcing, i.e. mutual cooperation is a (stable) social convention. In conclusion, some (possibly relatively small) degree of altruism may upset the noncooperative social convention; if moreover a large enough number of players behave (possibly just moderately) altruistically, an efficient, cooperative social convention could be eventually brought about.

Analogous arguments might be constructed for other games in which players' self–interest causes the emergence of inefficient social

[1] Specifically, the altruistic player will cooperate if $p > (\psi - \varsigma - (\lambda + \psi)w)/(\psi - \varsigma - \lambda)$; recall that the restrictions $\psi > \lambda + \varsigma$ and $w > \lambda/(\lambda + \psi)$ are needed here.

[2] At $w = (\psi - \varsigma)/(\lambda + \psi)$ she will cooperate even if she believes that the opponent will certainly defect.

conventions, as in the well known centipede game. These results suggest that altruism is not easily dismissable as an instance of irrational decision making and that, consequently, it might be an important motivating factor for players at least in certain situations. This possibility is somewhat backed by the available experimental evidence that shows that cooperative behaviour tends to be observed more often than it should be on the basis of the standard model of self–interested players and that cooperation is typically the 'default' initial choice (see e.g. Roth (1991)), as well as by the evidence that shows that equity considerations play a substantial role in individual bargaining decisions (see Ochs and Roth (1989)). In fact, recent experimental results by Frank, Gilovich and Regan (1993) suggest that subjects which have been exposed to microeconomics courses based on the notion of rational choice as self–interested maximizing behaviour tend to cooperate significantly less frequently than subjects who have not. This seems to indicate that the standard paradigm of self–interest may have a self–fulfilling nature: the exposed subjects are brought to perceive self–interest as a *normative* characterization of rational behaviour and to act accordingly. Thus, rather paradoxically, one could conclude that the takeover of the self–interest paradigm outside the narrow community of economists might induce an inefficient social convention at which cooperative deals become harder to enforce than it might otherwise be[3].

Once made clear that altruism may play an important role in the solution of some well known social dilemmas and that self–interest may work in the opposite direction, one has to explain why self–interest has traditionally become the dominant behavioural assumption on which the theory of rational decision making has been built. It is indeed true that a vast literature on altruism has developed nowadays (as a mere but representative example see Bernheim and Stark (1988), Stark (1989)). But what about the main message conveyed by such literature?

Some traits that are common to the various authors may be iden-

[3] Lattimore (1992) examines the results of a standard ultimatum bargaining game involving economists who behave according to the familiar self–interest paradigm and non–economists who value attitudes such as fairness and altruism. The (relatively) surprising finding is that economists make smaller expected monetary gains than non–economists.

tified:

i) altruism is represented as a *preexisting*, fixed 'stock';

ii) the analysis focuses on the allocative effects of the fixed stock;

iii) it turns out that if the *level* of altruism falls below a certain threshold, *negative* welfare effects emerge, in that, in order to prevent systematic exploitation, the altruist must undertake inefficient courses of action (the so called altruist's moral hazard);

iv) if the level of altruism falls above the threshold, welfare effects are instead positive;

v) *on the other hand*, given that high levels of altruism are not practically reachable, self–interested behaviour seems preferable: the welfare losses the economy must bear before a large enough stock of altruism is built are likely to be too large.

In conclusion, the main message of this literature seems to be that, all things considered, altruism is irrelevant and/or wasteful! It is our opinion that this conclusion heavily depends on an incorrect conceptualization of altruism and more specifically of its role in the complex motivational system that lies behind individual decisions. A first, peripheral remark is that altruism cannot be understood as a stock that is worn out by use, as it is the case with any scarce resource. Altruism is not a resource, it is a *virtue*. As a consequence, as Aristotle put it, its stock is likely to be *increased* by frequent use, rather than worn out. More importantly, the standard approach to altruism is flawed by a basic methodological error. It is assumed that the preference orderings of players do not change as the process of choice unfolds, i.e., it is implicitely postulated that behaviour and experience do not feed back on players' preference judgements. Now, while this assumption is certainly reasonable in comparative static analyses, it is much too restrictive when it comes to dynamics. A player's degree of altruism at a certain time is likely to be affected by his past history of choices and experiences; this in turn implies that his original preference judgements may be reshaped several times and in several different ways. A second, related point is that 'altruism' is a far too vague notion that calls for further qualification; in fact, one can consistently define several different *types* of altruism, all of which having peculiar allocative consequences. Therefore, once admitted that there are important feedbacks from the dynamic choice process to individual judgements, it must be recognized

that such judgements may evolve towards, or away from, different sorts of altruism, and that the actual law of motion may in principle heavily depend on the type of altruism that is currently actualized in the player's preference judgements.

In conclusion, a proper dynamical analysis of the allocative consequences of altruism must focus not only on the actual *level* of altruism that characterizes the economy, but also on the *type(s)*, keeping in mind that the time evolution of the stock(s) of altruism crucially depends on the distribution of the various types across the population. The dynamical mechanism just outlined lends itself quite naturally to an evolutionary dynamic treatment; in this perspective, the diffusion and the eventual takeover of a certain type of altruistic norm of behaviour (or of a certain distribution of types of altruism across the population) may be seen as the emergence of a social convention in the sense of Lewis and Sugden[4]. In this framework one can also easily model 'altruistically–driven' phenomena of transition between conventions, by studying if and how the appearance of some 'new' sort of altruism upsets the existing social convention laying the foundations of a new, different one, with the corresponding bearings in terms of social welfare.

The remainder of the paper is organized as follows. In section 2 we review some recent and some less recent literature on altruism and selfishness to elucidate the current state of the debate. In sections 3 and 4 we discuss in grater detail the issue of the relevance of altruistic motivations, as opposed to self–interest, in a well grounded theory of rational choice. Section 5 discusses the basic characteristics of our evolutionary approach. Sections 6 and 7 describe the reference model we use for our formal analysis. Section 8 discusses the dynamical structure of the model. Section 9 describes our results. Section 10 provides a short summary of our main findings.

2. A brief review of new and old literature.

[4] The need to model social conventions as the outcomes of density dependent dynamical evolutionary processes has already been pointed out by Bicchieri (1990).

The approach to rational choice that identifies self–interest as the sole motivating force behind individual decisions is still the dominant one in the current literature. Several justifications have been given for this choice. As Tullock (1976) puts it, "As a result of empirical research...the average human being is about 95 per cent selfish in the narrow sense of the term". On the other hand, Downs (1956), p. 29, observes that "In reality, men are not always selfish, even in politics...In every field, no account of human behaviour is complete without mention of such altruism...Nevertheless, general theories of social action always rely heavily on the self–interest axiom". Stigler (1981) explains economic behaviour as almost entirely self–interested. He contends that deviations from narrow self–interest are generally confined to one's "family, plus a close circle of associates". Mueller (1986), p. 14, argues that "the only assumption essential to a descriptive and predictive science of human behaviour is egoism".

Sen's 1977 celebrated paper (on "Rational fools") is the first major theoretical attack against the self–interest assumption. Sen argued that sympathy for other people and commitment to a principle produce two key departures from self–interest and that commitment which involves couterpreferential choice "drives a wedge between personal choice and personal welfare" while "much of traditional economic theory relies on the identity of the two" (p. 328). In 1978, Collard's "Altruism and economy" supported the thesis that "human beings are not entirely selfish, even in their economic dealings" (p. 3). Hirshman's "Against parsimony" in 1984 criticized the parsimonious postulate of a "self–interested, isolated individual" and argued for a new focus on changes in values (rather than 'wanton' tastes) and on noninstrumental action. Very recently, Kirzner (1990) has made the fine point that "Self–interest is indeed a central element, but this self–interest must...be understood with a certain subtlety. Properly understood, self–interest does not exclude altruistic motivation; it depends on purposefulness, but not on any selfishness of purpose. The point to be stressed is that it is one's *own* purposes which inspire one's actions and excite one's alertness. One's purposes may be altruistic or otherwise" (p. 39). In other words, it is, necessarily, the agent who makes the agent's *choices*, but the *interests* pursued need not be those of the agent. Thus, the familiar suggestion that every action is necessarily self–interested ("I

am pursuing *my* concern that others should be able to...") misses the point since it confuses the concept of choice as purposeful action with the concept of choice as self–interested action — a purpose is not the same thing as personal interest. As a general point, Frank (1988) argues that there is no need for economics to assume exclusive rationality, understood as complete freedom from emotion and passion. The validity of economics is therefore not threatened by pervasive real–world examples of passionate behaviour. Building on earlier work of Schelling and others, Frank shows how the practice of moral behaviour may not only be good for society, but it may also be materially beneficial to the practicing individuals themselves.

In a recent work, Sen (1987) marks the current state of the debate with the following, sweeping judgement: "Universal selfishness as *actuality* may well be false, but universal selfishness as a requirement of *rationality* is patently absurd" (p. 16).

Neither one should think that the tension between self–interest and altruism as motivating forces behind rational choice was not felt by the founding fathers of the modern orthodoxy. Steedman (1989) has a most illuminating discussion of Wicksteed's thought centered on the relations between rational choice theory, as commonly understood in economics, and a broader conception of the individual in society, who might have altruistic purposes, social and ethical ideals. Steedman's paper provokes important reflection on the relations between rationality and altruism: altruistic conduct, when it occurs, is fully consistent with rational economic conduct. Already toward the end of past century, Wicksteed (1933) pointed out that "The economic relation does not exclude from my mind every one but me, it potentially includes every one but you" (p. 174). To stress this point Wicksteed introduced the term 'non–tuism' to mean that, in an economic relation, A's lack of concern for the purposes of B (and vice versa) by no means entails that A acts for selfish motives. "The specific characteristic of an economic relation is not its 'egoism' but its non–tuism (p. 180). It follows from the above characterization of the nature of the economic relation that "The proposal to exclude 'benevolent' or 'altruistic' motives from consideration in the study of Economics is...wholly irrelevant and beside the mark" (p. 179).

In fact, by further widening our historical perspective, we discover

that even apart from the classical references to Mandeville and Smith, the issue of the role of altruism vs. selfishness in the working of complex economies and societies is far from being a recent discovery. Throughout history, economic philosophers were fascinated by the phenomenon that general welfare develops from egoism of merchants. It seems that the earliest writer who noticed this point was Anonymous Jamblichi, a Sophist and apparently an expert in economic ethics, who wrote around 450 B.C. He compared the wealth of the Greek city–states — the *poleis* — which did not have much gold, with the general Persian poverty that existed despite the immensurable treasure in gold possessed by the Great King. Anonymous Jamblichi solved this paradox by stating that it is the merchants' activity that caused Greek wealth. In turn, this activity of the Greek merchants was made possible by the manifold relations of trust and credit that existed as a legal framework both within and between the numerous city states in the Greek *koiné*. Thus, resulting from the egoism of the many, a general system of welfare developed (see Fikentscher (1988)).

The latin philosopher Seneca, in his *De Beneficiis*, says that a merchant selling corn does no benefit even if, in reality, he saves an individual or a whole city. Nobody of those who profit by his action is obliged to him, because he did not mean to help them, but had in mind his own interest (Seneca, VI, 14, 4). The sentence with which Seneca explains that whatever is done to make a profit is no benefit has a Smithian flavour: he brings advantage to other people in order to have his own (*ad alienum commodum pro suo veniunt*, IV, 13, 3). Therefore, merchants produce a 'good' result without having this purpose.

Much before the celebrated Mandeville's "Fable of the bees", in 1564 the German scholar Leonhard Fronsberger from Ulm published an essay, *Vom Lob deβ Eigen Nutzen* (*The praise of self–interest*), where he explained that the farmers, merchants and artisans are busy only in pursuit of their self–interest and in doing so create public order and well–being in the general interest (see Schultze (1987)).

Although very brief and of course quite incomplete, our survey demonstrates that the issue of the relevance of altruism in the motivational system of rational decision makers is far from being a marginal or ephemeral one; indeed, it may be regarded as one of the few issues that permeate the whole western intellectual history. Its somewhat marginal

role in the research agenda of economists looks therefore suspect and calls for a deep rethinking. To this purpose we devote the next two sections.

3. The selfishness assumption: ethical status and factual relevance.

The claim that self–interest alone motivates economic behaviour must be either tautological, as such deprived of any cognitive content, if self–interest can encompass any motive (altruism becomes a kind of sophisticated self–interest), or factually false, if self–interest means behaviour that consciously intends only self as the beneficiary (i.e. all behaviour is instrumental to one's self–interest and all actions are commensurate in terms of their impact on one's own welfare). It is therefore important to distinguish the two interpretations of the self–interest paradigm. The present section is devoted to the latter (the behavioural one) whereas the next is devoted to the former (the all–encompassing one).

The presumption that selfishness is a morally 'neutral' assumption, as Collard (1978) argues, is based on the view that, on a continuous scale ranging from envy, malevolence or hatred to sympathy, benevolence or love, self–interest lies in the middle. This position is however not defendable. Indeed, Collard seems here to confuse the notion of self–interest with its degree.

On the other hand, the fact that I neither hate nor love other people does surely imply that I am indifferent toward them, but it does not imply that I am selfish — as suggested by Collard - – since I might even hate myself. At times, the hatred–to–love feelings may represent the motivation for either egoism or altruism; but they do not lie on the same scale as egoism and altruism.

The point made here is not to deny the possibility or even the likelihood of man being selfish and therefore the legitimacy of a theory based on this assumption. What is rather argued is that the self–interest assumption is a non–neutral one on ethical grounds. The simplest way to see the point is the following. Start with the standard theory of rationality as utility maximization and add the assumption of self–interested

behaviour. Next notice that it is part of the meaning of the term 'rational' that a rational individual does not intentionally choose what is worse irrespective to the objective pursued by the individual. So rational *and* self–interested people never intentionally choose what is worse for themselves. Moreover, since according to the theory of rationality, intentional choice follows preference, rational and self–interested individuals never prefer what is worse for themselves to what is better. Finally, add a moral principle of minimal benevolence — a non controversial one indeed — according to which, other things being equal, it is a morally good thing if people are better off. Then it is, ceteris paribus, a morally good thing to satisfy an individual's preferences. As it can be seen, the self–interest assumption transforms rationality — which is not by itself a moral notion — into an ethically–laden concept. For this reason, we believe it is safe to claim for the altruistic assumption as much attention as that paid to its opposite — egoism.

It is of course beyond doubt that self–interest is an important motivator of individual choices and that not all individuals are concerned with the welfare of others. Indeed, self–interest explains *most* of human interaction in some contexts, typically in market–mediated transactions, and it plays *some* role in almost every context. We aim, however, at making thinking about self–interest more subtle: when people think about what they want, they think about more than just their narrow self–interest. Moral principles and the interests of others do have a weight in the definition of self–interest. Consequently, one can argue that people often take account of other individuals' interests when they decide what constitutes a 'benefit' they want to maximise; this extended set of motivations is of course perfectly consistent in principle with the idea that individuals choose rationally. What is in doubt is whether the *standard* approach to rational choice is able to capture the actual complexity of individual motivations.

It is a fact that people do co–operate more than the self–interest model (useful though it is) seems to predict. In their already cited paper, Frank, Gilovich and Regan (1993) shed light on one reason for this. Imagine a world in which people move from one prisoner's dilemma to the next (as it happens in real life situations). If people can choose their partners freely, and if honest types can spot each other in advance, co–operators will be able to interact selectively with each other — and

will therefore do better than cheats; this argument presents interesting analogies with the 'secret handshake' effect pointed out by Robson (1990) in an evolutionary game–theoretic context. Experiments have shown that people are surprisingly good at telling co–operators and cheats apart, even on the basis of what seems to be limited information. One can therefore argue that, on broad evolutionary grounds, narrowly self–interested behaviour might ultimately be self–defeating.

The possibility that narrowly self–interested behaviour may be 'selected away' from a population of rational decision makers at least under certain circumstances is an intriguing one — and is certainly at odds with most of the received wisdom about the relative survival possibilities of egoistically vs. altruistically motivated economic agents. One important avenue for future research would be to spell out the conditions under which self–interested and altruistic behaviours can coexist as opposed to those under which self–interest is likely to undermine altruism altogether.

In fact, it is hard to deny that a complex society is not likely to survive if *all* its members are governed by exclusively selfish motives. Although a counterfactual test of this statement is hard to implement or even to conceive, the available indirect experimental evidence tends to support it. For example, Dawes, van de Kraagt and Orbell (1988) have conducted a series of experiments over a decade, the results of which warrant the conclusion that the cooperation rate, in situations involving social dilemmas, can be enhanced in the absence of egoistic incentives.

To understand why these results do support our statement one must keep in mind that complex societies *cannot* survive without a certain amount of cooperation; see e.g. Argyle (1991). This of course does not rule out the possibility that there can be vital, complex societies that are characterized by an inefficiently low level of cooperation. A viable, complex economic system must balance between institutions that assume self–interest and those that assume the existence of broader sets of motivations. At the moment, institutions geared to self–interest probably have far more weight in liberal democracies than it is likely to be compatible with the long–run public good. (This is possibly a consequence of the self–interest bias that has characterized most of modern and contemporary thought about the structure and evolution of modern industrialized societies). Therefore, it makes sense to try to redress

this imbalance by revitalizing or creating institutions that foster a commitment to the common good. Yet efforts at revitalization will not be undertaken if it is common knowledge that such efforts will not work because of a widespread convinction that the only way to achieve efficiency is the promotion of selfish motives. In order to be able to design and implement institutions which rely on mixed sets of motives (selfish plus altruistic) it is necessary to acknowledge more fully the relevance and the potential of altruistic motivations and to give them a less marginal status in analytical and policy investigations.

4. Why shouldn't altruism be encompassed by the self–interest paradigm?.

We describe individuals as 'altruistic' when they feel and act as if the long–term welfare of others is taken as an end in itself; i.e. it is of relevance *independently* of its effects on their own welfare. If my concern with the welfare of others is merely an instrumental means for promoting my longer–term selfish ends and ceases once these selfish ends can be more easily pursued in some other way, I am an enlightened self–interested subject rather than an altruist subject.

One must distinguish between *degrees* and *varieties* of altruism. Every act falls somewhere on a line between extreme selfishness and extreme altruism, depending on the relative weight we give to our own interests and to the interests of others. Following Jencks (1979), we will distinguish *'complete'* altruism (counting our own good as neither more nor less important than anyone else's); *'partial'* altruism (taking *some* account of the interests of others), and *'extreme'* altruism (weighting other people's individual interests *more* heavily than our own). Jencks also distinguishes three sources of altruism: empathy, community, and morality.

Empathic altruism' derives from the fact that we 'identify' with people outside ourselves. We incorporate their interests into our subjective welfare function, so that their interests become our own.

'Communitarian altruism' involves identification with a collectivity rather than with specific individuals (group identity as solidarity).

'Moralistic altruism' involves the incorporation of external moral ideals into our sense of 'self'.

These three varieties of altruism can lead to quite different forms of behaviour. For instance, empathy usually induces us to promote the interests of some specific individual, whereas communitarian altruism tends to focus on larger groups, possibly imposing pain and sacrifices to individual subjects if this can help to promote the interest of the community. Therefore, these two forms of altruism can lead to conflicting prescriptions. Many moral dilemmas involve precisely such conflicts between varieties of altruism, rather than just conflicts between selfishness and altruism.

According to the tautological formulation of the self–interest paradigm, the above listed types of altruism can be all encompassed by the standard approach to rational choice once one properly specifies the subject's utility function. For example, consider an empathic altruist who cares about the well–being of another subject; call the two A and B, respectively. If we assume that the relative weights assigned by A to her own and to B's well beings are given by $1 - w$ and w, respectively, we can write a 'motivation function' for A

$$M_A = (1 - w)U_A + wU_B$$

where U_A and U_B are the utilities of the two subjects. At this point, one could claim that the assumption that A maximizes M_A correctly describes the motivations that lie behind A's choices. This claim, however, seems to miss the fundamental difference between the demand of consistency of choice and that of self–interest. Indeed, the former is a procedural demand without any substantive content as to what is being pursued, whereas the latter is a particularly substantive command as to what is being pursued without necessarily inducing a rigid procedural structure (see Sen (1987)). Therefore, the encompassing of altruism by the standard approach to rational choice would imply the recognition of the impossibility of an empirically based microeconomics, a consequence that very few economists would be willing to endorse (see Broome (1978)).

The 'reductionist' strategy just described need not be attacked on purely methodological grounds only. In the motivation function approach to altruism, the altruist subject A regards the utility of B as

instrumental; therefore, A will behave altruistically only to the point where *her* marginal benefit is equal to *her* marginal cost. From A's point of view, B's utility is something like a consumption good among others, apart from the obvious positive externality effect. The existence of a positive externality, however, gives rise to the well known free–riding problems. This implication seems to be markedly at odds with the large available evidence on individual donors contributing to large charities, which have only a marginal impact on the total income of the charity but a substantial opportunity cost; as a consequence, even putting methodological perplexities apart, this approach is not able to explain but a small part of observed altruism (see Sugden (1982)).

Another, more subtle encompassing strategy makes reference to the sociobiological literature. Becker (1976) and Tullock (1978), among others, borrow the inclusive fitness hypothesis that has been made popular by the work of Wilson (1975) and Dawkins (1976). According to this assumption, the natural selection process does not elicit the maximization of genes' own fitness but rather the maximization of the fitness of duplicate genes which happen to be carried by kin–related organisms. As a consequence, the mutual aid phenomena that are so commonly observed in nature are prompted by genetic programming aimed at increasing the probability of survival of the corresponding genes. In this perspective, altruism may be once again encompassed by a suitably expanded notion of self–interest.

The sociobiological framework allows for the modelling of the pursuit of beneficience and of the adherence to sub–optimal choice principles as an expanded set of tastes. This is the case because, as argued e.g. by Becker, the utility function of the agent *includes* as well the utility functions of other individuals. In this light, altruism, loyalty and trust are not really far–sighted strategies of self–interest, say, à la Axelrod (1984). Rather, they are tastes in themselves. It is important to stress that this reference to sociobiological categories does not mean that Becker makes a case for biological reductionism; indeed, he imports only the superficial make up of the inclusive fitness hypothesis, not its biological content; more specifically, Becker explains the allocation of endowments between self–interest and other–interest by means of the standard individual utility maximization procedure rather than by means of the maximization of a genetic fitness criterion.

Hirschleifer (1977a,b, 1982(1987)) also regards altruism and loyalty as tastes, i.e. not as merely contingent outcomes of educated self–interested calculations. However, he regards such tastes as genetically guided: "Economic man's behaviour is constrained by *inbuilt* emotions and tastes. While these no doubt contain accidental elements, they are not completely arbitrary. What tastes sweet to us is mainly what serves our own interest and our 'irrational' or 'unselfish' drives have largely met the evolutionary test of enabling us the better to compete via group membership" (Hirschleifer (1982(1987)), p. 263)[5].

Frank (1988) seems also to regard 'unselfish' tastes to be genetically programmed. Frank blends standard analytical tools with a Darwinian selection mechanism in order to explain the prominence of 'irrational' sentiments. He treats analytically the tastes for conscience, revenge, altruism as part of a continuous utility function.

Whether such tastes are chosen for their own sake à la Becker or as a consequence of specific genetical programming à la Hirschleifer and Frank, one cannot help noting that sociobiologically based explanations of altruism provide at best a partial picture of the phenomenon; in particular, none of these authors seems to be willing to embrace a fully reductionist position. On the other hand, the link between natural selection at the genetic level and economic selection at the behavioral level still remains obscure at the current state of knowledge. In lack of a general explanatory principle that is able to connect and elucidate the interplay of the various selection mechanisms, it is hard to claim that sociobiological ideas may provide the basis for an all–encompassing theory of rational choice.

In order to provide a fully convincing account of altruism as a motivating force behind individual rational choice, it is therefore necessary to embed altruistic concerns in a more fully specified theory of extended rationality. The remaining sections of this paper are aimed at moving some steps in this direction.

---

[5] The competition via group membership advanced by Hirschleifer amounts to the advocacy of evolution via group–selection, as opposed to via individual–selection. Group selection is opposed by ultra–Darwinists like Dawkins but has found some credit in the more recent literature; see e.g. Eshel (1972).

5. An alternative evolutionary approach to altruism.

The fact that evolutionary explanations of altruism based on the sociobiological principle of inclusive fitness maximization at the genetic level do not provide us with a well founded theory of extended rationality does not rule out the possibility of building an alternative evolutionary approach based on different principles. The recent developments of the debate around the sociobiological paradigm seem to hint that, after all, radical genetic reductionism is a somewhat far fetched position as far as the explanation of human conduct is concerned (see e.g. Gallino (1980)). As extensively argued e.g. by Harris (1989), genetic factors are only able to account for a narrow minority of human customs and behaviours, the majority of which seem instead to be the outcome of a much more teleologically oriented process of cultural evolution whose action is often not traceable to any specific characteristic of the genetic substratum.

It seems therefore reasonable, as a first approximation, to study the evolutionary mechanisms that account for the selection of *behaviours* abstracting from the action of genetic factors. Unlike natural selection, cultural selection is dominated by the transmission of individual experiences from one generation to another; culture itself is after all nothing but a sophisticated mechanism aimed at the storing of past experiences and at the selective prompting of new ones.

The long run dynamics of cultural evolution processes is very complicated and requires a unified approach which melts together the insights coming from the various social sciences such as economics, sociology, anthropology, social psychology, history and so on. It is not excluded that in the future the need of such a widening of perspective will be acknowledged in order to arrive at a truly satisfactory approach to rational choice. For the moment, we confine ourselves to a much less ambitious objective and limit our attention to the short(–medium) term dynamics of cultural evolution. We will assume that, in the short term perspective we focus on, the evolution of behaviors is essentially driven by the imitation of relatively more rewarding behaviours. We do not mean to deny that learning can occur through much more complex modes including structural inference, active research or sophisticated adaptive rules (see e.g. Selten (1991) for a comprehensive survey on

this topic); our aim, however, is to make the model as simple as possible in order to elucidate its basic features especially as far as the modelling of altruistic behaviour is concerned; for a general discussion of the methodological issues raised by this model see Sacco (1993).

We have so far criticized the existing approaches to the modelling of altruistic behaviour by arguing that a) they fail to account for the dynamic, history–dependent feedback effects from experience to tastes through which altruism is reinforced within, or wiped out from, the individual's motivational system; b) there exist different varieties of altruism which can induce quite different behaviours as well as quite different 'motivational dynamics' in the above specified sense. We will argue that these two crucial features may be captured by means of a simple dynamic model of cultural evolution. Instead of trying to build a general theory, we will concentrate on a specific example, a contribution game with two donors and one receiver. We will distinguish the various sorts of altruism in terms of the respective preference orderings on the outcomes of the game and will show under what conditions they are selected (or selected away) when matched to different types of altruism (or of egoism). We will moreover be able to extract a few stylized facts which are likely to be exportable to some degree to a more general context. It is important, however, to stress that we mean this exercise as a preliminary step that calls for extensive further exploration in later work.

It is worth remarking that the taxonomy of altruist types we introduce below is not the Jencks' taxonomy discussed in section 4. The relationship between our and the Jencks' taxonomy is not immediate because the two sit at different levels of specificity; Jencks classifies altruism according to its possible *motivations*, whereas our classification criterion is based on the *preference orderings* that are induced by the various types. As a consequence, each of our altruist types may in principle be compatible with more different motivations in the sense of Jencks.

Our taxonomy is meant as an original contribution that partly draws on the previous literature and has been designed for the specific model we have in mind, although its validity is probably not limited to this narrow context.

6. The contribution game and player types.

In this section we introduce the following contribution game: two parties (I and II) have to decide whether to contribute a sum of money to a third party (P) who is in a state of need. To simplify matters, we assume that, in order to survive, P needs to receive one dollar; I and II have to decide whether to contribute one dollar, half a dollar or nothing. The 'payoff' the contributors get depends on the amounts they contribute (respectively) and on the total amount collected by P. Given our assumptions, the outcome space for the contribution game contains nine elements; we index the outcome space by means of couples $(\cdot,\cdot)$, where the first entry indicates I' s contribution whereas the second indicates II' s contribution.

In order to classify various possible types of players, the first basic distinction to be introduced is between altruists and egoists. In the present context, altruistic players are those whose 'payoff' depends not only on the amount contributed but also on the amount collected by P. Egoists, on the other hand, do not care for P' s well being; their main concern is to minimize their expenditure. Within the class of altruistic players, various types of altruism are characterized by the corresponding preference orderings $\succ$ over the outcome space. In particular, we distinguish four different types of altruists[6]. Take for example the point of view of player I. Then we say that I is a *Subsidiary Altruist* (SA) if

$$(0,1/2) \succ_I (1/2,0)$$

$$(0,1) \succ_I (1/2,1/2) \succ_I (1,0)$$

$$(1,0) \succ_I (0,0)$$

$$(1/2,0) \succ_I (0,0)$$

[6] Miller (1989) considers as a special instance of a contribution game the case concerning the institution of a welfare state. The basic question he raises is: "What must be true of people's altruistic concern for others if they would prefer to see a welfare state in existence rather than relying on private charitable schemes?" (p. 100).

(Here and in the following, we will restrict only to strict subsets of possible outcome comparisons, i.e. those which characterize specifically the particular type of player we are considering).

In other words, a Subsidiary Altruist is an individual I that cares for P' s well being but prefers that, if possible, the contribution rests on the shoulders of the other player, II. On the other hand, if I knew that II is not willing to contribute, she would contribute something rather than nothing.

I is instead said to be a *Reciprocal Altruist* (RA) if

$$(0,0) \succ_I (1/2,0) \succ_I (1,0)$$

$$(1/2,1/2) \succ_I (0,0)$$

$$(1/2,1/2) \succ_I (0,1)$$

A Reciprocal Altruist is thus an individual whose aspiration is that of having himself and the other player both contributing the same amount. If I now knew that II is not going to contribute, he would contribute nothing. On the other hand, I would prefer to contribute something if he knew that II is going to contribute something. In other words, a Reciprocal Altruist is someone who wants that the responsibility for contribution is equally divided among all subjects; if this is not the case, he prefers giving up contributing rather than carrying the burden of contribution alone[7].

We say that I is a *Kantian Altruist* (KA) if

$$(1/2,0) \succ_I (0,0)$$

$$(1/2,0) \succ_I (1,0)$$

$$(1/2,1/2) \succ_I (0,1)$$

[7] Trivers (1971) calls 'reciprocal altruism' an ongoing pattern of interaction in which I help you, then you help me and so on. Reciprocal Altruism is a matter of 'enlightened self–interest' in the sense clarified by Axelrod (1984). It is interesting to note that Reciprocal Altruism seems to depend on trust and that in turn trust probably depends, at least to some extent, on empathy.

A Kantian Altruist is someone who is ready to contribute independently of the other player's decision. On the other hand, if I knew that II is not going to contribute, she would contribute one half rather than one, i.e., she would contribute her part of an equal sharing of the burden of contribution, in spite of the fact that this contribution alone is not enough to warrant P' s well–being. Moreover, I prefers a situation where both players contribute something rather than a situation where all the burden is on the shoulders of the other player[8].

Finally, I is a *Superkantian Altruist* (SKA) if

$$(1,0) \succ_I (1/2,0) \succ_I (0,0)$$
$$(1/2,1/2) \succ_I (0,1)$$

The main feature that distinguishes a Superkantian Altruist from a Kantian Altruist is that, if she knew that the other player is not going to contribute, she would be ready to carry all the burden of contribution alone rather than merely her part of an equal sharing of the burden. In other words, whereas a KA is mainly concerned with the fulfilment of her moral duty to contribute, a SKA' s main concern is P' s actual well being. Coming to egoist types of players, we say that I is a *Pure Egoist* (PE) if

$$(0,1) \succ_I (1/2,1) \succ_I (1,1)$$
$$(0,1/2) \succ_I (1/2,1/2) \succ_I (1,1/2)$$
$$(0,0) \succ_I (1/2,0) \succ_I (1,0)$$

In other words, a Pure Egoist is someone who is neither concerned with P' s well being nor is willing to contribute anything for any reason.

---

[8] Laffont (1975) introduces and formalizes the notion of Kantian behaviour in a model with externalities and public goods. He shows that efficiency will result if everybody acts as a Kantian. The Kantian contribution is defined as the maximum amount that each individual would perceive as morally right to contribute in each group he belongs to. Sugden (1984), building upon Laffont, presents a theory based on a notion of reciprocity and Kantian behaviour. Other related works are those of Collard (1978, 1991).

Alternatively, we say that I is an *Egoist with Regret* (ER) if

$$(0,1) \succ_I (1/2,1) \succ_I (1,1)$$
$$(0,1/2) \succ_I (1/2,1/2) \succ_I (1,1/2)$$
$$(0,0) \succ_I (1/2,0) \succ_I (1,0)$$
$$(0,0) \succ_I (0,1/2) \succ_I (0,1)$$

That is, an Egoist with Regret is someone who is not interested in P' s well being but nevertheless is disturbed by the fact that someone else is willing to contribute, the more so the higher the amount of the contribution. Notice that this does not imply any evidence of concern for P' s well being: I' s 'payoff' is higher the less P receives; moreover, she will not contribute anything anyway.

The previous discussion may be resumed by Figure 1 below:

[Insert Figure 1 here]

In order to proceed with the formal analysis, we have to introduce some more structure. In order to make 'payoffs' of the various types of players commensurable, we will identify for each a 'bliss outcome' (not necessarily unique) to which we will assign a 'payoff' level equal to one. As to outcomes which are less preferred w.r.t. the bliss one, we will assign 'payoff' levels equal to $1/2$, 0, $-1/2$ or -1, according to cases, thus completing the above introduced preference orderings and choosing a specific numerical representation. There are of course many possible completions and numerical representations that are compatible with the preference judgements listed above; the one chosen does not seem to entail, however, a relevant loss of generality in any specific sense.

To avoid unnecessary complications, we assume that players' (negative) 'payoffs' at outcomes in which there is overcontribution are the same irrespectively of the specific type. Indicating the 'payoff' function for I as $\pi(\cdot,\cdot)$, we therefore postulate that:

$$\pi(1,1) = -1 \quad \pi(1,1/2) = -1 \quad \pi(1/2,1) = -1/2$$

The numerical representation of players' preferences on the remaining six outcomes according to the respective types are the following:

a) Subsidiary Altruist:

$$\pi_{S\,A}(0,1) = 1 \quad \pi_{S\,A}(0,1/2) = 1/2 \quad \pi_{S\,A}(0,0) = -1$$
$$\pi_{S\,A}(1/2,0) = 0 \quad \pi_{S\,A}(1/2,1/2) = 1/2 \quad \pi_{S\,A}(1,0) = 0$$

b) Reciprocal Altruist:

$$\pi_{R\,A}(0,1) = -1/2 \quad \pi_{R\,A}(0,1/2) = 0 \quad \pi_{R\,A}(0,0) = 1/2$$
$$\pi_{R\,A}(1/2,0) = -1/2 \quad \pi_{R\,A}(1/2,1/2) = 1 \quad \pi_{R\,A}(1,0) = -1$$

c) Kantian Altruist:

$$\pi_{K\,A}(0,1) = -1/2 \quad \pi_{K\,A}(0,1/2) = 0 \quad \pi_{K\,A}(0,0) = 0$$
$$\pi_{K\,A}(1/2,0) = 1/2 \quad \pi_{K\,A}(1/2,1/2) = 1 \quad \pi_{K\,A}(1,0) = 0$$

d) Superkantian Altruist:

$$\pi_{S\,K\,A}(0,1) = -1 \quad \pi_{S\,K\,A}(0,1/2) = -1/2 \quad \pi_{S\,K\,A}(0,0) = -1/2$$
$$\pi_{S\,K\,A}(1/2,0) = 0 \quad \pi_{S\,K\,A}(1/2,1/2) = 1 \quad \pi_{S\,K\,A}(1,0) = 1/2$$

e) Pure Egoist:

$$\pi_{P\,E}(0,1) = 1 \quad \pi_{P\,E}(0,1/2) = 1 \quad \pi_{P\,E}(0,0) = 1$$
$$\pi_{P\,E}(1/2,0) = -1/2 \quad \pi_{P\,E}(1/2,1/2) = -1/2 \quad \pi_{P\,E}(1,0) = -1$$

f) Egoist with Regret:

$$\pi_{E\,R}(0,1) = 0 \quad \pi_{E\,R}(0,1/2) = 1/2 \quad \pi_{E\,R}(0,0) = 1$$
$$\pi_{E\,R}(1/2,0) = -1/2 \quad \pi_{E\,R}(1/2,1/2) = -1/2 \quad \pi_{E\,R}(1,0) = -1$$

Notice that this numerical representation for ER implies that I' s negative 'payoff' caused by the observation of the (positive) contribution of others is substantially less than the negative 'payoff' caused by her own contributing.

The formal description of the contribution game is thus complete. In the next section we derive best reply strategies for each type of player depending on the opponent' s type.

7. Best replies.

SA vs. SKA. Playing against a SKA, a SA will never choose to contribute 1; the reason is simple: since a SKA will contribute something anyway, contributing 1 for a SA becomes a dominated strategy. Thus the choice for SA is restricted to contributing 0 or 1/2. In the first case, KA' s best response is contributing 1; in the second, KA' s best response is contributing 1/2. On the other hand, provided that KA contributes 1 (1/2), SA' s best response is to contribute 0 (1/2 or 0). We have therefore two possible Nash equilibria: $(SA = 0, SKA = 1)$; $(SA = 1/2, SKA = 1/2)$. On the other hand, there is a strong incentive for SA to play 0 anyway, since even provided that SKA intends to play 1/2, SA gets in this case the same 'payoff' by playing 0 or 1/2, i.e., $SA = 1/2$ is a weak best response to $SKA = 1/2$. On the other hand, if SKA plays 1, SA is strictly better. We can therefore conclude that, provided that a SA and a SKA meet, the natural outcome to predict is $(SA = 0, SKA = 1)$.

SA vs. KA. As before, a SA knows that, playing against a KA, the latter will always choose to contribute something. Therefore, $SA = 1$ is ruled out. Now, irrespectively of whether $SA = 1/2$ or $SA = 0$, KA' s best response is playing 1/2. SA is indifferent between these two options: provided that KA plays 1/2, SA gets 1/2 anyway. This choice, however, does make a difference for KA: if SA plays 1/2, by playing 1/2 KA gets 1, whereas if SA plays 0, KA gets only 1/2. To sum up, in this case there are two Nash equilibria: $(SA = 0, KA = 1/2)$, $(SA = 1/2, KA = 1/2)$. Neither of them can be ruled out, so both outcomes are possible.

SA vs. RA. In this case, SA knows that a RA will never play 1. If RA plays 0, SA is indifferent between playing 1 and playing 1/2. However, provided that SA plays 1/2, RA prefers to play 1/2 as well. Now, if RA plays 1/2, SA is again indifferent between playing 0 and

playing 1/2. Therefore, the only reason why a SA may be induced to play 1 is the belief that RA will play 0. In this case, however, SA will be indifferent between playing 1/2 and playing 1. But if RA is not going to play 0, playing 1/2 is definitely better for SA since she will be strictly better off provided that RA plays 1/2. As to RA, if SA plays 1 a (weak) best response for RA is to play 0; if instead SA plays 1/2, RA will surely play 1/2. Finally, notice that SA will never play 0 in equilibrium. In conclusion, there are two Nash equilibria: $(SA = 1/2, RA = 1/2)$ and $(SA = 1, RA = 0)$. As in the case SA vs. SKA, the second Nash equilibrium may be ruled out. We are therefore left with a unique prediction.

SA vs. PE. In this case, PE unambiguously plays 0. SA is then indifferent between playing 1/2 or 1. We therefore obtain the two Nash equilibria $(SA = 1, PE = 0)$, $(SA = 1/2, PE = 0)$, none of which can be ruled out.

SA vs. ER. In this case also ER unambiguously plays 0, and SA is therefore indifferent between 1/2 and 1. This choice, however, now makes a difference for ER; if SA plays 1, ER gets 0, whereas if SA plays 1/2, ER gets 1/2. (Notice the analogy with the case SA vs. KA).

RA vs. KA or vs. SKA. These cases are relatively straightforward. A RA knows that a KA will play 1/2 anyway, so she will play 1/2 as well; the same reasoning holds a fortiori for the SKA. Therefore, in both cases the natural outcome to predict is $(RA = 1/2, KA = 1/2)$, resp. $(RA = 1/2, SKA = 1/2)$. (In the case of a SKA, $(RA = 0, SKA = 1)$ would in principle be a Nash equilibrium as well but, since $RA = 0$ is a weak best response, it can be ruled out by the same line of reasoning set out above).

RA vs. PE or ER. Here also the analysis is straightforward. Both PE and ER will always play 0. Consequently, RA will surely play 0 as well.

KA vs. SKA. Provided that the KA will play 1/2 anyway, the SKA will play 1/2 as well. Once again the natural outcome to predict is $(KA = 1/2, SKA = 1/2)$. (Here too there exists another Nash equilibrium which can be ruled out by the usual argument, namely $(KA = 0, SKA = 1)$).

PE vs. KA or SKA. Provided that $PE = 0$, KA will surely play $1/2$, whereas SKA will surely play 1.

ER vs. KA or SKA. Also ER will surely play 0. However, if facing a KA (who will play $1/2$), ER gets $1/2$, whereas if facing a SKA (who will play 1), ER gets 0.

ER vs. PE. Clearly, both will play 0 reaching their bliss outcome.

The only cases left are now those where players face opponents of the same type.

SA vs. SA. When two SA meet, it is easy to check that three Nash equilibria are possible: i) one of the two contributes 1 whereas the other doesn' t contribute; ii) both contribute $1/2$; iii) one of the two contributes $1/2$ whereas the other doesn' t contribute. At the last two equilibria, both players get the same payoff; at the first, the one who leaves the burden of contribution on the shoulders of the other is better off.

RE vs. RE, SKA vs. SKA, KA vs. KA. In all three cases, the natural outcome to predict is clearly that at which each player contributes $1/2$.

PE vs. PE, ER vs. ER. In both cases, the natural outcome to predict is clearly that at which each player contributes nothing.

8. Evolutionary dynamics: basic assumptions.

In this section we move one step further to analyze the following model. We start from a continuous population in which players of two different types initially coexist. Time is continuous. Players are randomly matched to play the contribution game discussed earlier. Players' types are perfectly observable, as well as the actual distribution of types within the population. When matched to an opponent of a certain type, a player acts according to the theory of strategic rationality set out in the previous section, i.e. they play their best response strategy to the opponent' s type. It is moreover assumed that, at each given time t, a negligible fraction of players is prompted to revise their strategy choice,

according to the following criterion: a player decides to change her type if and only if the strategy that characterizes the other type yields a higher expected 'payoff', i.e. behaviours that look more rewarding are imitated at the expense of others. In other words, the timing of players' decisions is asynchronous and players are boundedly rational in the sense that their strategy choices may (temporarily) not be best responses to the observed distribution of types. Notice however that the actual best response strategy depends on the distribution of types itself; this means that a given strategy may work in some situations but not in others. In spite of players' bounded rationality, if a stationary distribution of types is ever reached, it should have the property that players eventually learn the corresponding best response strategy.

The set of assumptions just introduced is summarized by the following dynamic model, known as replicator dynamics [see e.g. Hofbauer and Sygmund (1988)]:

$$\dot{q} = q(1-q)\Delta\tilde{\pi}(q) \tag{1}$$

Here q denotes the proportion of players of type α (say) in the population; the proportion of players of type β is clearly $1-q$. $\Delta\tilde{\pi}(q)$ is the (expected) 'payoff' differential between type α and type β players, which depends on q. $q = 0$ and $q = 1$ are always stationary points for (1), that is, if the population is entirely made up of players of a same type, it will stay homogeneous as long as no exogenous perturbation occurs. Finally, a type' s proportion increases if and only if it performs better than the other, as required above. It is easy to check that equation (1) is equivalent to postulating that $\dot{q}$ is positive if and only if players of type α earn a 'payoff' that is higher than the population average 'payoff' $q\tilde{\pi}_{\alpha}(q) + (1-q)\tilde{\pi}_{\beta}(q)$.

In our context, players' strategy choices concern whether to be altruist or egoist, and of which type. This clearly implies that players' altruistic (egoistic) attitudes are to be seen as the object of a purposeful choice rather than as a basic, immutable feature of their personality on which they have no control. Players are assumed to be ready to choose a certain altruistic (egoistic) attitude whenever the observation of other individuals reveals that people behaving that way look 'happier' than others. Of course, players' 'happiness' (viz., 'payoffs') has to be perfectly

observable; one can for example think that the level of 'happiness' may be unambiguously inferred from the intensity of an individual' s 'smile' (or 'frown'). This is the mechanism through which players' experience is assumed to feed back on the respective motivational systems. Moreover, notice that the diffusion of a certain kind of altruistic (egoistic) attitude within the population may be easily read as the emergence of a social convention in the sense of Lewis (1969): given that people tend to behave in a certain way, each individual finds convenient to behave in the same way and expects the same holds for all others. In fact, it is possible to show that all asymptotically stable states of the evolutionary dynamics (1) (i.e., all possible social conventions in the sense of Lewis) are evolutionarily stable Nash equilibria[9].

In the following section we will study under what conditions such social conventions emerge starting from given initial distribution of types. In this paper we limit our attention to two–type populations; however, there is nothing that prevents the extension of our analysis to higher–dimensional cases. This endeavour, which is clearly of great interest, is left for future research.

9. Selection of conventions: The relative 'fitness' of altruistic and egoistic attitudes.

The order with which the various possible cases are analyzed in this section is different from the one followed earlier. In both cases, it has been chosen for expositional convenience; no confusion should arise from this.

SKA vs. ER. We begin by analyzing the case of an initial population made up of SKA and ER players (briefly, a SKA–ER population). Denote by q the proportion of SKA players. Their expected 'payoff' is then given by

[9] This result no longer holds when more than two types of players interact, as far as the evolutionary stability part of the statement is concerned; see Hofbauer and Sygmund (1988).

$$(2) \quad \tilde{\pi}_{S\,K\,A}(q) = q\pi_{S\,K\,A}(1/2, 1/2) + (1-q)\pi_{S\,K\,A}(1,0) = \frac{1}{2}q + \frac{1}{2}$$

Accordingly, the expected 'payoff' for ER players is

$$(3) \quad \tilde{\pi}_{E\,R}(q) = (1-q)\pi_{E\,R}(0,0) + q\pi_{E\,R}(0,1) = 1-q$$

One therefore has that

$$(4) \quad \Delta\tilde{\pi}(q) = \frac{3}{2}q - \frac{1}{2}$$

from which it follows that $\Delta\tilde{\pi}(q) = 0$ for $\hat{q} = 1/3$. More specifically, for $q > \hat{q}$ one has that $\Delta\tilde{\pi}(q) > 0$, whereas for $q < \hat{q}$, $\Delta\tilde{\pi}(q) < 0$. It is then easily checked that the following holds:

Proposition 1. *Consider a SKA–ER population. If the initial proportion q_0 of SKA players is larger than one third, then all players will eventually become SKA. If conversely q_0 is less than one third, all players will eventually become ER.*

What Proposition 1 says is that, in order to observe a stationary population entirely made up of SKA players when the alternative for players is to be of the ER type, a large enough initial proportion of SKA players is necessary [see Figure 2 below]. The critical threshold which qualifies the 'large enough' clause clearly depends on the numerical specification of 'payoffs'. Assume for example[10] that instead of $\pi_{S\,K\,A}(1,0) = 1/2$ one has $\pi_{S\,K\,A}(1,0) \equiv \phi$, where $\phi \in [0,1]$. Then one has that $\tilde{\pi}_{S\,K\,A}(q) = \phi + (1-\phi)q$ and $\hat{q} = (1-\phi)/(2-\phi)$. As $\phi \to 0$, i.e. SKA players get little pleasure from carrying the whole burden of the contribution alone, $\hat{q} \to 1/2$, i.e., the minimal initial proportion of SKA players needed to bring the SKA convention about increases. On

[10] Here and in the following we assume that, when one or more 'payoffs' are allowed to vary, other 'payoffs' vary accordingly if necessary in order to ensure that the selected actions are still best responses.

the other hand, as $\phi \to 1$, i.e., as SKA players tend to find their bliss in carrying the burden alone, $\hat{q} \to 0$, i.e., the population is eventually all made up of SKA players as long as a positive initial proportion of SKA players, however tiny, is present in the population.

[Insert Figure 2 about here]

Notice that, however little the pleasure SKA players get from carrying the burden alone (as long as $\phi > 0$), if SKA players are initially the majority within the population, everybody will be a SKA eventually.

PE vs. ER. We now pass to a population where players are all egoists, although of different types. It is clear that, in this case, both types of players attain the same 'payoffs' whatever the initial distribution of types since they will always play exactly the same way. As a consequence, the initial distribution of types q_0 will not be altered as time unfolds: players have nothing to learn from experience. Every possible q_0 is therefore a social convention here. In other words, in this case (1) gives rise to a neutral dynamics.

SKA vs. PE. Denoting again by q the proportion of SKA players, one has

$$(4) \quad \tilde{\pi}_{SKA}(q) = q\pi_{SKA}(1/2, 1/2) + (1-q)\pi_{SKA}(1,0) = \frac{1}{2} + \frac{1}{2}q$$

On the other hand, it is easily checked that $\tilde{\pi}_{PE}(q) = 1$, i.e., PE players reach their maximum attainable 'payoff' anyway. Consequently, we have

Proposition 2. *Consider a SKA–PE population. Unless the population is initially entirely made of SKA players (i.e., $q_0 = 1$), all players will eventually become PE.*

Notice the difference between the results of Proposition 1 and those of Proposition 2. The fact that egoist players may regret, i.e. may be disturbed by the generosity of Superkantian players, gives Superkantian Altruism a (pretty large) chance to survive in equilibrium. On the other hand, Pure Egoists are not affected at all by the behaviour of

their opponents, however altruistic it is. Therefore, they will be always happier than SKA altruists and Pure Egoism will prevail [see Figure 3]. If however in the limit SKA players attain a 'payoff' of $\phi \to 1$ when carrying the whole burden of the contribution, they end up as happy as PE players. In other words, Superkantian Altruism is not wiped out by Pure Egoism only in its 'pure' form, i.e. when SKA players reach their bliss when carrying the whole burden of the contribution, valuing altruism *per se*. In this latter case, a neutral dynamics results once again, that is, the initial distribution of types is preserved: both Pure Egoists and altruists keep their mind, learning nothing new from experience.

[Insert Figure 3 here]

KA vs. ER. Denoting by q the proportion of KA players and computing 'payoffs' by the usual procedure, we have

$$(5) \qquad \tilde{\pi}_{K\,A}(q) = q\pi_{K\,A}(1/2,1/2) + (1-q)\pi_{K\,A}(1/2,0) = \frac{1}{2}q + \frac{1}{2}$$

$$(6) \qquad \tilde{\pi}_{E\,R}(q) = (1-q)\pi_{E\,R}(0,0) + q\pi_{E\,R}(0,1/2)$$

from which it follows

$$(7) \qquad \Delta\tilde{\pi}(q) = q - \frac{1}{2}$$

As in the case of SKA vs. ER, it is easily checked that $\Delta\tilde{\pi}(q) > 0$ for $q > \hat{q} = 1/2$ and $\Delta\tilde{\pi}(q) < 0$ for $q < \hat{q}$. Therefore:

Proposition 3. *Consider a KA–ER population. If the initial proportion q_0 of KA players is larger than one half, then all players will eventually become KA. If conversely q_0 is less than one half, all players will eventually become ER.*

Notice the difference w.r.t. the case SKA vs. ER. Being less generous than those of SKA players, the choices of KA players are less

disturbing for ERs. Therefore, the minimum initial proportion of KA that is necessary to bring Kantian Altruism about eventually is larger than that for Superkantian Altruism. Clearly, as KA players get more satisfaction from their partial contribution to P' s well being (in spite of its inadequacy), i.e. as $\pi_{K\,A}(1/2,0) \equiv \alpha \to 1$, such constraint is loosened, i.e. $\hat{q}$ decreases. Unlike the case of SKA players, it is improper here to speak of pure altruism, in that in this case the KA player is simply not concerned at all by the fact that the needs of P have not been completely fulfilled, i.e. she reaches her bliss despite this. Rather than a pure altruist, such player seems rather a pure philistine.

More generally, assume now that $\pi_{K\,A}(1/2,0) \equiv \alpha$, $\pi_{E\,R}(0,1/2) \equiv \beta$, $\alpha,\beta \in [0,1]$. β clearly measures the extent to which ER players are disturbed by the altruism of KA players; the closer β to 1, the less ER players are disturbed. In this case, $\hat{q} = (1-\alpha)/(2-\beta-\alpha)$. As α and β both tend to zero, $\hat{q} \to 1/2$. On the other hand, if $\alpha \to 0$ whereas $\beta \to 1$, $\hat{\mu} \to 1$ (i.e., all players are eventually ER for $q_0 \neq 1$); if instead $\alpha \to 1$ whereas $\beta \to 0$, $\hat{\mu} \to 0$ (i.e. all players are eventually KA for $q_0 \neq 0$). Finally, if both α and β tend to 1, the limit value of $\hat{q}$ depends on the relative speeds of α, β.

KA vs. PE. Denote by q the proportion of KA players. It is easy to check that $\tilde{\pi}_{P\,E}(q) = 1$, as above. As to KA, we have

$$\tilde{\pi}_{K\,A}(q) = q\pi_{K\,A}(1/2,1/2) + (1-q)\pi_{K\,A}(1/2,0) \tag{8}$$

Therefore,

$$\Delta\tilde{\pi}(q) = \frac{1}{2}q - \frac{1}{2} \tag{9}$$

from which it follows

Proposition 4. *Consider a KA–PE population. Unless the population is initially entirely made of SKA players (i.e., $q_0 = 1$), all players will eventually become PE.*

Analogously to the case SKA vs. PE, for $\pi_{K\,A}(1/2,0) \equiv \delta \to 1$ one obtains a neutral dynamics which preserves the initial distribution

q_0 whatever it is. As observed above, this sort of behaviour cannot be interpreted as an instance of pure altruism.

KA vs. SKA. This case is very easy to be dealt with, since starting from any initial distribution of types, players' behaviour is the same irrespectively of the actual type of the opponent with which they are matched, provided that the Nash equilibrium $(KA = 1/2, SKA = 1/2)$ at which both types reach their bliss is played. If instead the other Nash equilibrium $(KA = 0, SKA = 1)$ is played, all players are eventually SKA since $\pi_{KA}(0,1) < 0 < \pi_{SKA}(1,0)$. This latter possibility is however somewhat unlikely for the reasons explained earlier.

RA vs. PE or ER. Here also the analysis is very simple, since RA players always choose to contribute nothing against egoist types of players. As a consequence, RA players achieve a low level of 'payoffs' whenever they meet egoist players, whereas the latter always get their maximum attainable 'payoffs' (no matter whether they meet RA players or players of their same type). Thus, unless the initial population is entirely made of RA players, everybody will eventually be egoist here. Notice in particular that Reciprocal Altruism is the only sort of altruism that is not disturbing for ER players.

RA vs. KA or SKA. When RA players meet KA or SKA players, each subject chooses to contribute one half, no matter her type. As a consequence, players always reach their bliss indipendently of the identity of the opponents with which they are matched. Once again we therefore have a neutral dynamics that preserves the initial distribution of types q_0, in which players have nothing to learn from esperience.

Egoists players with probabilistic regret. We have so far only considered egoist players which either feel no regret or always regret when meeting players who choose to contribute. We now consider the case of egoist players who regret with a fixed probability $\theta \in (0,1)$ provided that their opponents choose to contribute, or of egoist players who regret with a probability θq, increasing in the proportion q of altruistic players (with θ still belonging to the interior of the unit interval). It can be checked that, in the former case, the critical threshold $\hat{q}$ for a KA–ER(θ) population, beyond which all players eventually become KA, is equal to $\hat{q} = 1/(1+\theta)$. For a SKA–ER(θ) population, the threshold

is now equal to $\hat{q} = 1/(1+2\theta)$. In the latter case, the critical threshold for a population becomes $\hat{q} = (1/2\theta)[\sqrt{2\theta+1/4} - 1/2]$. Notice that $(1/2\theta)[\sqrt{2\theta+1/4} - 1/2] < 1/(1+2\theta)$, i.e., the survival chances of egoist behaviour are enhanced when they regret with probability θq since as q is close to zero (i.e. as there are few SKA players around) egoist players are less likely to regret. More generally, the case of egoist players who regret with a certain probability (no matter whether fixed or density–dependent) is intermediate between the case of PE and that of ER players; the survival chances of egoist behaviour vary accordingly.

SA vs. KA. We have observed earlier that when two SA players meet there are three possible Nash equilibria: $(0,1)$, $(0,1/2)$, $(1/2,1/2)$. In order to determine how two SA players choose their strategy when meeting each other, we assume that they play a correlated equilibrium that prescribes to play $1/2$ with probability $1-\gamma$ and 0 and 1 with probabilities $\gamma/2$ each; this is equivalent to a prescription to play the equilibrium (1/2,1/2) with probability $1-\gamma$ and the equilibrium (0,1) with probability γ, with the assignment of roles randomized with equal probabilities. Notice that the equilibrium (0,1/2) has been ruled out here since at this equilibrium players get at most as much as they get at the equilibrium (0,1) (and strictly less when they are not prescribed to contribute).

At the correlated equilibrium (c), players expect to get

$$(10) \qquad \frac{\gamma}{2}\pi_{SA}(0,1) + \frac{\gamma}{2}\pi_{SA}(1,0) + (1-\gamma)\pi_{SA}(1/2,1/2) = \frac{1}{2}$$

We have moreover seen that, when playing against KA players, SA players have two possible best responses: playing 0 or playing 1/2. Denote by q the proportion of SA players and assume that they choose to play 0. Then we have

$$(11) \qquad \tilde{\pi}_{SA}(q) = q\pi_{SA}(c) + (1-q)\pi_{SA}(0,1/2) = \frac{1}{2}$$

$$(12) \qquad \tilde{\pi}_{KA}(q) = (1-q)\pi_{KA}(1/2,1/2) + q\pi_{KA}(1/2,0) = 1 - \frac{1}{2}q$$

from which it follows

$$\Delta\pi(q) = \frac{1}{2}q - \frac{1}{2} \tag{13}$$

Following the usual line of reasoning, it is easy to check

Proposition 5. *Consider a KA–SA population. Unless the population is initially entirely made of SA players (i.e., $q_0 = 1$), all players will eventually become KA.*

The rationale behind Proposition 5 is clear: when two SA players meet, they cannot simultaneously reach their bliss; on the other hand, when two KA players meet, they always reach their bliss. On the other hand, when meeting an opponent of the other type, both types of players get the same 'payoff'. Therefore, Kantian Altruism spreads over the population because of its superiority w.r.t. Subsidiary Altruism in cases where players of the *same* type meet.

Consider now the more general case where $\pi_{S\,A}(0, 1/2) = \epsilon$, $\pi_{K\,A}(1/2, 0) = \eta$. Then one can check that there exists a critical threshold $\hat{q} = (1 - \epsilon)/(3/2 - \epsilon - \eta)$ beyond which all players will eventually be SA. $\hat{q}$ lies in the interior of the unit interval if $\eta < 1/2$. In this case, Subsidiary Altruism has a chance to survive provided that the initial proportion of SA players is large enough; the less KA players are happy to carry the burden of contribution alone, the less this constraint is binding.

If on the other hand SA players choose to play 1/2, the 'payoff' for the two types are now independent of the actual distribution of types and are given by $\tilde{\pi}_{S\,A}(q) = 1/2$, $\tilde{\pi}_{K\,A}(q) = 1$. We are therefore led to an interesting conclusion: although playing 0 or 1/2 is indifferent from the point of view of SA players facing a KA player, the latter choice makes KA players definitely happier and thus leads to the eventual disappearance of Subsidiary Altruism (independently of the actual 'payoff' specification for SA players). In other words, even if playing 0 does not give SA players a direct advantage in terms of 'payoffs', it gives them an *evolutionary* advantage (in the sense of the survival possibilities of Subsidiary Altruism). More generally, if SA players randomize between playing 0 and playing 1/2 with probabilities ρ and $1 - \rho$, respectively,

it turns out that $\hat{q}$ lies in the interior of the unit interval if $\rho > 1/2$ and $\eta < 1 - (1/2\rho)$, a more binding condition than $\eta < 1/2$. Clearly, the larger ρ, the more the survival possibilities of Subsidiary Altruism.

SA vs. SKA. Once again there are two possible Nash equilibria when a SA player meets a SKA player: both playing 1/2 or the SKA player carrying the whole burden of the contribution. We consider first this latter equilibrium and denote by q the proportion of SA players. In this case, it is easy to check that

$$\tilde{\pi}_{SKA}(q) = \tilde{\pi}_{SA}(q) = 1 - \frac{1}{2}q \tag{14}$$

Since both types get exactly the same 'payoffs' independently of the actual distribution of players' types, it follows that $\Delta\tilde{\pi}(q) \equiv 0$, i.e., the dynamics preserves the initial distribution of types. KA players have therefore an evolutionary advantage w.r.t. SKAs when playing against SAs: unlike SKAs, they contribute only their fair part of the overall burden, thus making SA players relatively worse off. Notice moreover that, in this case, SA players (expect to) get 1/2 when playing against themselves and 1 when playing against SKAs; SKA players, on the other hand, get 1 when playing against themselves and 1/2 when playing against SA. An immediate corollary is therefore that if $\pi_{SKA}(1,0) < 1/2$, all players will eventually be SA, whereas if $\pi_{SKA}(1,0) > 1/2$, all players will eventually be SKA. In other words, if SKA tend to be pure altruists, Superkantian Altruism spreads over in a SKA–SA population.

As to the other Nash equilibrium where both SKA and SA players contribute one half, it is easily checked that now $\tilde{\pi}_{SKA}(q) = 1$. As a consequence, all players will eventually be SKA now. Once again, then, although the two possible options for SA players are indifferent in terms of 'payoffs', they are definitely not indifferent in terms of the survival chances of Subsidiary Altruism.

Finally, if SA players randomize between playing 0 and playing 1/2, with probabilities σ and $1 - \sigma$, respectively, it can be checked that all players are eventually SKA unless $\pi_{SKA}(1,0) \equiv \chi < 1/2$ and σ is large enough, more precisely $\sigma > 1/2(1 - \chi)$.

SA vs. PE or ER. For a SA–PE population, it is very easy to check that all players will eventually be PE. As to a SA–ER population, there

are once again two possible Nash equilibria when a SA player meets a ER player: the one at which the SA player contributes one half and the one at which she contributes 1. Clearly, ER players are more disturbed when SA players contribute 1; more specifically, in this latter case both the SA and the ER player get a 'payoff' of 0. However, even so all players will eventually be ER since ER players perform better than SA players when playing against an opponent of their same type.

SA vs. RA. In this case, all players will eventually be RA since $\tilde{\pi}_{S\,A}(q) = 1/2$ whereas $\tilde{\pi}_{R\,A}(q) = 1$. This is due to the fact that, when playing against RA players, SA can do no better than playing 1/2. Thus RA players here do better than SAs both when playing against opponents of their same type and against opponents of the other type.

10. Conclusion: a few 'stylized facts' and directions for future research.

We conclude our formal analysis with a list of the most important 'stylized facts' that can be extracted from the conclusions reached in the previous section.

a) Pure Egoism can never be wiped out provided that it is initially present in the initial distribution of types. There are some forms of altruism, namely Kantian and Superkantian Altruism, that can survive against Pure Egoism even if only in extreme cases (i.e. when players are 'pure altruists' or 'pure philistines').

b) If egoist players are disturbed by the altruism of their opponents, egoistic behaviour can be wiped out under certain conditions. The sort of altruism that performs better against Egoism with Regret is Superkantian Altruism. Some sorts of altruism, like Subsidiary or Reciprocal Altruism, are instead systematically wiped out by egoistic behaviour of any sort.

c) There are some forms of altruism, like Kantian or Superkantian Altruism, that although performing relatively well against egoistic behaviour can be wiped out by other forms of altruism, specifically Subsidiary Altruism, at least under certain conditions.

d) There are many cases in which players cannot learn from experience and the initial distribution of players' types reproduces itself perpetually. This happens whenever the types that make up the population behave homogeneously (but can also happen when types do not behave homogeneously, as in the SA vs. SKA case).

Our analysis may be extended in several different directions. One could study higher–dimensional population dynamics with more than two behavioural types. Also, one could interestingly extend our results to more general game–theoretic contexts; it is not excluded that further relevant variants of selfish and altruistic behaviour may then emerge. At a more fundamental level, our results, in line with some other recent game–theoretic literature, call for a deep rethinking of the methodological individualism postulates on which the standard theory is built. It is not asked to abandon methodological individualism, but rather to go *beyond* it: we have argued that the core elements on which purposeful individual choices are based (i.e. preferences) may themselves evolve as a consequence of repeated social interaction. Therefore, a true understanding of the nature and structure of social institutions calls for a redistribution of emphasis away from the individual level and toward the *relational* level. Economic phenomena have a primary social dimension. Individual behaviours are *embedded* in a preexisting network of social relations which cannot be thought as a mere constraint; rather, they are one of the driving factors that prompt individual goals and motivations (see Granovetter (1985)). People's aspirations are deeply conditioned by the conventional wisdom about what makes life worth living. What we need is a new organizing principle, that we could provisionally call methodological relationism, to throw light on the interplay between individual behaviours and the architecture of society. Several interesting contributions which point in this direction have appeared recently; see for example von Foerster (1984) and Alexander and Giesen (1987). This is of course a very complex and far–reaching set of issues, which the current literature has just begun to explore; as a consequence, there isn' t yet a satisfactory theoretical framework that allows to address them in full generality. Despite this, the evolutionary approach to altruism set out in this paper seems to us a promising point of departure for further thought in this vein.

To conclude, the main message we would like to convey by this

paper is that the attempt of relaxing the assumption of selfishness in favour of a notion of altruism appears to us as the most promising research strategy in coping with several present–day economic problems. After all, it is not at all clear why rationality should not involve the intelligent pursuit of *all* one's goals and values, properly weighted, rather than sticking only to a particular class of goals, i.e. self–interested ones[11]. Indeed, by not taking goals seriously, the standard concept of rationality has offered an impoverished view of human behaviour. An extended view of rationality, one which recognizes people's ability to think about their goals, emerges as a superior alternative to the standard, instrumental view.

[11] For a general discussion of this question see Zamagni (1991), where morality is seen as an element in an extended preference function.