EXTENDED RATIONALITY, ALTRUISM AND THE JUSTIFICATION OF MORAL RULES

by Stefano Zamagni
Department of Economics, Bologna
July 1991

1. Introduction

Ever since A. Smith reasoned about men's Moral Sentiments, as well as the nature and causes of the Wealth of Nations, the relation between ethics and economics has been an issue of scholarly interest. Current and received contributions on the issue seem to fall roughly into two categories. The first includes contributions which discuss the potential role of ethical judgements in economic theory itself. The second category comprises contributions which discuss ethics as an aspect of human behaviour, i.e. as part of the empirical reality studied by economists.

Attention, here, is focused on the second category. The purpose is to analyse how ethics can be incorporated as a potential explanatory variable into economic theory.

Attempt to account for a dispositional variable like morality in economic explanations seem to be at odds with the standard economic classification of explanatory variables into either preferences or constraints. Given this system of reference, introducing morality into economic explanations as a variable would seem to require us either to classify morality as a subjective, intrapersonal preference-variable or as an objective, external constraint-variable. Since classifying morality as an external

constraint is ruled out by the fact that morality is defined as an intrapersonal, dispositional variable, the only alternative seems to be to view morality as a preference-variable. This view is implicit in the dichotomy between morality and self-interest which is quite common in discussions of the relationship between ethics and economics. (See Stigler, 1982).

The proponents of such a dichotomy seem to suppose that some modification of the standard economic model of self-interested behaviour is needed in order to account for morality. There seem to exist two view. One conceptualises morality as a motivating force which is beyond and outside of any interest calculation or as a factor that is outside a person's preference function and is not captured by the notion of rational choice. The other view conceptualises morality as part of a person's preference function, i.e. it considers human interests to be broader than self-interest and to include something like an interest in being a moral person.

The first view is problematic at the conceptual level. If the principle of rational, self-interested behaviour is meant to imply that individuals choose the alternative they expect to serve their interests best, it is difficult to see how morality can be conceptualized as a factor that makes individuals choose something different from what they think is in their best interest. One would have to assume that man has "two natures", an interest-oriented one and a moral one and that sometimes the one and sometimes the other determines his behavioural choices. Without a more general theory which would integrate the two at some higher level and specify the conditions under which the one or the other of the "two natures" will prevail, explanations of observed behaviour in terms of such a model would be totally arbitrary. Until now, there is no such general theory which could actually be compared with the model of rational self-interested choice-although many interesting attempts are under way. (Sen (1977); Hirschman (1984); Elster (1986); Steedman and Krause (1986)).

The second view, which sees morality as an element in an extended preference function, cannot be accused of the conceptual difficulty of the first view, even though economists following a Beckerian framework (Becker, 1981) would prefer explanations in terms of constraints over those in terms of preferences and changes in preferences. (A similar perspective is the one elaborated by Gauthier (1986) in his attempt to show "why an individual, reasoning from non-moral premises, would accept the constraints of morality on his choices" (p.5)).

In what follows, I will explore the potential of this second view with respect to a specific problem: the making sense of altruistic behaviour within a rational model of conduct. My general target is showing that we cannot fully describe the economic consequences of altruism if we do not understand the formation and the justification of altruism.

2. Rational choice and moral principles

The relation between rational choice and moral principles is an unresolved problem in the history of ethics. The problem is posed at the very beginning of the history of philosophizing about the justification of moral rules, in Plato's Republic. Glaucon acknowledges that agreement on certain moral rules can serve the mutual interests of those who are parties to the agreement. But he goes on to argue, by appeal to the story of the ring of Gyges, that the ways of morality cannot be fully justified in terms of rational self-interest. If one did in fact possess a ring that made one invisible, it would be in one's own interest to use it to secure one's own ends, even though this would clearly mean violating the usual moral constraints. Gyges used the ring in just this way: He seduced the queen, killed the king, and seized the throne for himself.

What the story of the ring of Gyges appears to do is drive a

wedge between the concept of rational choice and that of choice that respects the usual kinds of moral constraints. Moral rules, in short, exhibit the features of public goods and are thus subject, on the standard account of rationality, to the problem of free riding. The rational injunction, then, with respect to moral rules is highly conditional: Follow these rules unless it is in your interest to violate them.

Moral rules, however, are supposed to be constraints on an individual's choice that cannot be set to one side by considerations of what would serve the rational interests of that individual. From this it is usually taken to follow that what is needed is a different sort of justification of moral rules altogether - one that proceeds by reference to something other than the rational interests of the agent whose behavior is thereby to be constrained. For some, like Bentham and Mill, appeal is made to a generalized principle of interest - doing what serves the interests of the greater number; for others, like Kant, appeal is made to some allegedly non-interest-based principle (e.g., the categorical imperative).

As remarked by McClennen (1990), Glaucon fails to note, however, the symmetry of the situation: All do, in fact, possess the ring of Gyges, since for each there are occasions on which some moral rule can be violated with (relative) impunity. Moreover, it is plausible to suppose that all would do better if all were to accept the moral constraints as binding than if all were to defect whenever this could be done with impunity. This can be the case even though it is also true that holding the behavior of all others as constant, each is better off defecting. All of this, of course, suggests that an agreement to abide unconditionally by at least certain moral rules would be mutually advantageous, even though, of course, such an agreement would be unstable, as judged from the perspective of the traditional way of thinking about non-strictly-competitive games.

However, I believe the time has come to reconsider the tra-

ditional solution concept for non-strictly-competitive games. One need not suppose that a gap exists between rational pursuit of interest and commitment to moral principles. That gap exists only for those who persist in conceptualizing rational choice in terms of the principle that each should maximize with respect to his own interests. On the alternative view of extended rationality, according to which it is at least prima facie rational for each individual to do his part to maintain a mutually advantageous arrangement, the gap disappears.

3. Altruism within ends-rationality

The apparent existence of altruism has been seen as a strong challenge to the familiar form of ends-rationality.

To recall, the key elements in any analysis of rationality would include the individual's beliefs, preferences and decisions. In general, rationality is usually defined in terms of properties of one or more of these three elements or their interrelationships. A first interpretation of rationality can be approached in either of two ways. In one approach, consistency of actions is the hallmark of rationality, regardless of the beliefs and preferences of the decision maker. The second approach focuses on the interrelationship between beliefs and preferences within the process of choice. Rationality is defined as efficiency in pursuing given ends in the context of given beliefs. We know that the two approaches are fully equivalent under standard assumptions. Following Max Weber, we may call the first interpretation of rationality as consistency/efficiency means-rationality. A second interpretation of rationality relates exclusively to the individual's preferences. We will call it ends-rationality. A third interpretation of rationality is as a condition on the beliefs which carries no direct implication for either preferences

or decisions. According to belief-rationality, an agent is rational to the extent that his internal model of the world is accurate and well-informed.

Useless to say, the three interpretations are neither mutually esclusive nor collectively exhaustive. They provide a basic framework which is of value in characterizing the view of rationality embedded in economic theory, a view often referred to as utility maximisation. Actually, the latter expression is used in two distinct senses which relate to two different interpretations of rationality. In one sense, utility-maximisation is a purely formal specification of individual motivation having no substantive implications for the content of individual preferences. In the so-called "present-aim theory" (Parfit, 1984) rationality is taken to be the efficient pursuit of whatever aims one has at the moment of deliberation and action. As Sen (1977) notes: "if you are consistent, then no matter whether you are a single minded egoist or a raving altruist or a class conscious militant, you will appear to be maximising your own utility in this enchanted world of definitions" (p.323). If means-rationality is the only type of rationality that is allowed, the status of utility-maximisation is entirely restricted to the "as if" methodological position. (Some views of ethics are argued to depend only on means-rationality; for example, Rawls (1971) avoids any particular notion of ends-rationality for his constituents in the original position).

Ends-rationality in economics provides us with the second sense of utility maximisation. Perhaps the most frequently quoted statement of the particular form of ends-rationality is provided by Edgeworth: "the first principle of economics is that every agent is actuated only by self-interest" (p.16). Here, then, utility maximisation takes on a substantive form and it becomes a special case of the more general self-interest theory. It says an act is rational if it efficiently promotes the interests of the person who performs it. A person who does not cheat when he could

get away with cheating is judged, by the self-interest standard, to have acted irrationally. (On the other hand, a person who refrains from cheating because of guilt feelings would be considered rational by the "present-aim theory" standard, even if there were no possibility that cheating could have been detected). Ironically, however, the self-interest version allows us to say that a self-interested person might very well want to be motivated by moral sentiments (provided that their presence is recognizable by others). He may even take purposeful steps to enhance the likelihood that he will develop these sentiments. (He may join a church, for example). Once he acquires the sentiments, of course, the behaviour they provoke will still be officially categorized as irrational by the self-interest standard. Under the present interpretation one must assume that all "interests" are commensurable into a single dimension - utility. So self-interest and commensurability are both required components of utility-maximisation in its interpretation as a form of ends-rationality. In the following I will always refer to this interpretation. (I will not touch upon the uses in economic theory of the notion of belief-rationality. Hargreaves Heap (1989) provides a thorough survey).

Coming to the issue I started from in this section, the first question to be addressed is the possibility that altruistic behaviour can be incorporated within the standard view of ends-rationality. We will see that this is not a satisfactory response to the challenge of altruism. An alternative model is that of <u>extended</u> rationality: own utility is not the sole motivator of individual action and there exists an internal tension between self-interest and a number of distinct motivators or commitments.

There are three distinct lines of argument which attempt to reconcile altruistic behaviour with ends-rationality. The first is based on the economics of externalities and simply claims that the donor's utility is positively related to the utility of the recipient. The crucial point here is that the recipient is viewed

entirely instrumentally by the donor so that an individual will behave altruistically only to the point where his marginal utility is equal to his marginal cost. The recipient's utility is simply a consumption good like any other seen from the point of view of the donor.

A severe difficulty confronting this line of argument lies in its inability to overcome the free-rider problem. Much altruistic behaviour involves individual donors contributing to large charities. The self-interest model of altruism suggests that each such potential donor faces the classic public good problem in which his own contribution increases his costs markedly while increasing the total income of the charity (and therefore his own utility) only marginally. The prediction of free-riding do not appear to fit the facts: this approach explains no more than a small part of observed altruism (Sudgen, 1982).

A second possibility for claiming that altruism is, at base, self-interested is to argue that the act of giving itself produces utility. One version of this argument, which seems to fit the analysis of repeated games, is the claim that altruistic behaviour may build a valuable reputation for the donor. However, this type of argument is thin. Behaving altruistically can build a favourable reputation only if others are unaware of the underlying self-interest. For it to be possible to masquerade as an altruist, it must be the case that genuine altruists exist and that the public cannot distinguish between real and bogus altruism. The possibility of bogus altruism depends on the existence of real altruism and the latter is still unexplained. While this second line of argument does escape from the free-rider problem, it cannot provide a fully satisfactory explanation of altruism. Coleman (1983) provides an interesting analysis - albeit a very limited one - of what he calls the rationality of the zealot as opposed to the rationality of the free-rider.

The third possible method of incorporating altruism into rational egoism involves a strategic or evolutionary appeal to

self-interest (Akerlof, 1980; Margolis, 1982). The general idea is that while altruism may appear to be un-self-interested in the short-run, its long-term benefits - including the benefit of living in a society of altruists - may dominate these short-run costs even in the egoist's private calculus. Again, the problem here is the temptation to free-ride, since the best position for each individual is that of an egoist in an altruistic group. For altruism to be an evolutionary stable strategy in a society of individuals motivated by enlightened self-interest, some means of over-coming this free-rider problem has to be endogenously provided.

The above discussion brings us to the central paradox which lies at the heart of all attempts to explain altruism within self-interest: even if it is possible to argue that altruism will achieve increased personal utility for all, it is not generally possible to find a pure individualistic means of realising this outcome which is available to egoists.

4. Degrees of altruism and efficiency

The genuine altruist - as opposed to the self-interested altruist - provides a possible escape from the prisoner's dilemma (PD). Under pure egoism each agent places total weight on his own utility pay-off, completely ignoring the pay-off of the other agent. A genuine altruist may place some positive weight on another's utility even though it does not affect his own utility. Such an altruist may then be motivated by the weighted sum of his own utility and that of the other agent, where w is the weight placed on the other's utility and (1-w) the weight placed on own utility.

To illustrate this situation, table I represent a standard PD. Let's identify agent 1 as the altruist and let $w=\frac{1}{2}$. It follows that the cooperative strategy is a dominant strategy for a-

gent 1. Indeed, this result follows whenever $w>5/12$ (The required condition is: $0(1-w)+12(w)>5(1-w)+5w$, hence $w>5/12$)

Table I		C	D
	C	(10,10)	(0,12)
	D	(12,0)	(5,5)

Table II		C	D
	C	(10,10)	(4,12)
	D	(8,0)	(5,5)

The general point here is that there is a level of altruism- interpreted as a value of w - sufficient to ensure that the altruist selects the C strategy. For values of w below this critical level, altruism by itself is not sufficient to provide the basis for cooperation. However, any w such that $2/12<w<5/12$ transforms the game of Table I into an assurance game which is distinguished from a PD by the fact that free-riding is not the most attractive option (The condition for this is: $10(1-w)+10(w)>12(1-w)+0>5(1-w)+5w>0+12w$, hence $2/12<w<5/12$. Indeed, an assurance game is defined as one in which the outcomes are ranked by agent 1 as follows: $(C,C)>(D,C)>(D,D)>(C,D)$. See Sen, 1967).

Table II indicates the transformed matrix of pay-offs for the case $w=1/3$. Of course, it is no longer possible to interpret the pay-offs to agent 1 as own utilities, but they still represent his choice ordering. From the point of view of altruistic agent 1, no dominant strategy now exists since his best strategy depends upon agent 2's strategy. How then should 1 behave if he cannot wait for 2 to reveal his strategy? If he could be assured of 2's choice, his problem is trivial, he simply follows suit; but in the absence of such complete assurance, it is often suggested that he should defect (Cfr. Elster, 1984, p.21). However, as argued by Collard (1978) assurance need not be complete in order for 1 to choose the cooperative strategy. Agent 1 will cooperate if $p>5/3-4w$, where p is 1's subjective estimate of the probability that 2 will cooperate (The condition is that the expected pay-off from cooperation must exceed the expected pay-off from defec-

tion: $10(1-w)p+10wp+0+12w(1-p)>12(1-w)p+0+5(1-w)(1-p)+5w(1-p))$.

So the greater is agent 1's altruism, the less assurance he requires to justify his own cooperation. (With $w=5/12$ he will cooperate even if he believes that 2 will certainly defect).

The above indicates that the degree of altruism is of considerable significance in determining the nature of its effect. (Bernheim and Stark (1988) demonstrate that sufficiently high levels of altruism almost always lead to efficient resource allocation). At low levels ($w<2/12$) altruism does not materially affect the situation, the altruist still confronts a PD and his dominant strategy - even accounting for his altruism - is to compete. At moderate levels ($2/12<w<5/12$) altruism transforms PD to an assurance game but it is not sufficient to guarantee cooperative behaviour by the altruist - a degree of trust or assurance is still required. Finally, at high levels ($w>5/12$), altruism is sufficient to ensure that cooperation is the altruist's dominant strategy.

So, extending the individual's rationality beyond self-interest provides a potential escape from PD and, more generally, changes the nature of the tension between ratioanlity and ethics.

Of course, it is possible to argue that such extended rationality is formally equivalent to the reassertion of egoism. We could write a "motivation function" for agent 1 of the form:

$$M_1 = (1-w)U_1+wU_2$$

and then say that 1 maximises M_1. This would amount to incorporate altruism within egoism by the first argument discussed in the previous section. However, this argument misses the fundamental point that M_1 is not a utility funtion and that utility is not the sole motivator. A point to note here is that the recognition of motivations other than own utility breaks the link between preference and behaviour, so that it can no longer be argued that action reveals preference (Broome, 1978).

Clearly, the example above is subject to free-rider problems in any more general, n-person, setting. The question of whether the notion of extented rationality might effectively overcome this problem in the setting of altruism remains an open one. Some forms of extented rationality do seem to escape this trap - e.g. the "fair share" argument of Margolis (1982); the reciprocity argument of Sugden (1984), or the Kantian moral view discussed by Laffont (1975), Collard (1978). The point here, as in the assurance game framework, is that pure altruism will only rarely and in extreme cases be sufficient to make any significant difference to individual behaviour.

Only when altruistic concern combines with particular views regarding individual obligations and duties, or particular evolutions of the behaviour of others, will altruism be individually effective. In other words, altruistic concern has to be embedded in a more fully specified theory of extented rationality.

5. The nature of altruism: a) needs versus preference altruism

A promising line of attack on the specification of such extended rationality begins with the attempt to provide the individual decision-maker with a more detailed internal structure. To this end, let's consider, by way of example - albeit a relevant one -, the problem that arises if we assume that people are altruistically concerned about the welfare of others.

The question I want to ask is: what must be true of people's altruistic concern for others if they would prefer to see a welfare state in existence (i.e. an institution with the following three features: first, it provides benefits to everyone regardless of whether they have contributed to the cost of providing them. Second, it provides specific benefits which are seen as meeting needs. Third, the institution is funded by mandatory taxation) rather than relying on private charitable schemes as sug-

gested by libertarians? To answer this we need to look at the nature of altruism itself. What does it mean to be concerned about the welfare of others?

(I should make it clear that I do not regard the altruism argument as the only possible foundation for the welfare state. One could argue that those in need simply have rights to the resources which will meet their needs, independently of what other in their society believe or prefer. Libertarians deny the existence of rights to positive provision. So, my purpose here is to show that an argument can be constructed on premises that both sides can endorse. It will have a conditional character: if people have altruistic concerns of this type, then they will consent to institutions of such and such a form. Since libertarians profess to be neutral as to which conceptions of the good people should hold, they cannot escape the force of such an argument).

Let's consider the meaning of altruism. There is a number of possible ways in which the interests of others can enter the practical deliberations of the person in question.

A first contrast has to do with the way in which the interests of the others are interpreted. Does the altruist give the preferences of the others canonical status or does he employ some other notion of interests such as meeting the needs of the others as he defines them? Collard (1978) describes altruism of the latter kind as "meddlesome". A meddlesome altruist - or as I would call him a needs altruist - will want to try to prevent the recipient of his aid from converting it into a form that is preferred by the latter but valued less by the donor.

On the other hand, a preference altruist is somebody who interprets altruism on the lines of a Humean notion of sympathy. The other person's welfare matters to us because his happiness strikes a resonant chord in our frame: we take delight in the other's pleasure and sorrow in his pain.

Can the notion of ends-rationality cope with altruistic preferences? Or does it only work with selfish individuals?

There is a famous proposition, dating back to Edgeworth, which asserts that it makes no difference whether selfish or unselfish preferences are assumed. The intuition behind this conjecture is that individuals are in a game of exchange with each other, driving hard bargains, and it does not matter what motivates those bargains: it is the fact of hard bargaining and exchange which is crucial to the results.

The proposition can be examined with the aid of a simple example (Green, 1971). Suppose there are two individuals A and B and two commodities x_1 and x_2:

$$U_A = U_A\ (x_{A1},\ x_{A2},\ x_{B1});\ U_B = U_B\ (x_{B1},\ x_{B2},\ x_{A1})$$

In other words, the consumption of the first commodity by each individual affects the utility of the other. Nevertheless, the utility function is still well behaved in the usual sense. The equation:

$$\frac{dU_A/dx_{A2}}{dU_B/dx_{B2}} = \frac{dU_A/dx_{A1} - dU_A/dx_{B1}}{dU_B/dx_{B1} - dU_B/dx_{A1}}$$

gives the condition which must be satisfied for A's utility maximisation in these circumstances and a similar expression can be derived for B. There are two ways this can be derived: the simplest sets up the individual's utility maximisation problem in the usual way and embodies the constraints that the allocation of each commodity sums to its supply and that the utility of the other exceeds some arbitrary value. When the equation above and its analogue for B hold, then there is a Pareto optimum and the allocations satisfying this condition constitute the contract curve for the economy.

When the interdependence between individuals is non-meddlesome, meaning that the allocation of x_1 to B only influen-

ces the utility of A through the effect that x_1 has on the utility of B, then:

$$\frac{dU_A}{dx_{B1}} = \frac{dU_A}{dU_B}\,\frac{dU_B}{dx_{B1}} \quad \text{holds.}$$

Substituting this into the previous equation produces:

$$\frac{dU_A/dx_{A2}}{dU_A/dx_{A1}} = \frac{dU_B/dx_{B2}}{dU_B/dx_{B1}}$$

and, if the same non-meddlesome attribute holds for B, then the Edgeworthian proposition has been proved. The last equation defines the same contract curve as would exist in an economy where the utility functions of A and B were selfish. This is an important result. It means that a competitive equilibrium is still (potentially) Pareto optimal even though the preferences of the individuals are interdependent. The point is that, when individuals maximise their utility with respect to the competitive price vector, the last condition above will be satisfied whether preferences are interdependent or not. Under competitive conditions, each individual can only choose his own consumption of x_1 and x_2; they have no control over the consumption of x_1 by the other individual and so utility maximisation produces the same condition as when preferences are purely self-regarding. So the contract curve is achieved even though preferences are "other-regarding" - but non-meddlesome. Even though competitive markets offer no mechanism to take account of "other-regarding" preferences a (potentially) Pareto-optimal allocation is nevertheless produced with respect to those preferences.

Therefore, the inclusion of non-meddlesome or preference al-

truistic behaviour would not seem to be precluded by the concept of ends rationality.

The contrast between preference altruism and needs altruism has an obvious bearing on the case for a welfare state. Preference altruists will want to provide the objects of their concern with readily convertible resources, enabling them to reach their highest level of (self-defined) welfare (the simplest form of provision being cash redistribution). Thus they will be attracted to negative income tax schemes and the like. Needs altruists will want to ensure that certain specific needs are met and will favour provision in kind.

However, there are special considerations which may lead preference altruists some way in this direction as well. If we suspect that people are liable to make choices that are bad from the point of view of their long-run welfare, preference altruists too may favour provision in kind.

6. b) Calculating, reciprocal, Kantian altruism

I turn now to a second contrast between varieties of altruism, this one cross-cutting the preference/needs contrast. It presupposes a context in which there are a number of possible donors. Each potential donor has a personal interest in not making a contribution; other things being equal, he would like to keep his resources to spend on himself. On the other hand, if he were the only possible donor, he would give up to a certain amount. How will people behave?

There are two altruists A and B, equally endowed, facing C who is in need to the extent of 1 unit of resources. A and B, in isolation, would be willing to transfer 1 unit to C. Assume each can choose to give: 0, $\frac{1}{2}$ or 1. There are then 9 possible outcomes. Considering just A, there are potentially as many forms of altruism as there are rank orderings of these nine outcomes. Rea-

listically, however, we can narrow the range somewhat. (We can disregard the three outcomes $(\frac{1}{2}, 1)$, $(1, \frac{1}{2})$, $(1,1)$ in which C ends up with more resources then he needs). I shall confine my attention to four possibilities.

The first I shall call the calculating altruist. He is a person who wants to see C helped, but as far as possible by someone else. This means that:

$(0, 1) > (\frac{1}{2}, \frac{1}{2}) > (1, 0)$

$(0, \frac{1}{2}) > (\frac{1}{2}, 0)$, where ">" stands for "preferred to".

Depending on the strength of A's altruism we may either have $(1,0) > (0, \frac{1}{2})$ or $(0, \frac{1}{2}) > (1, 0)$.

This altruist is a calculating altruist because of the way in which his behaviour depends on his assessment of how other people will behave. If he expects to be able to get away without contributing, he will. There is, however, no reason to doubt his concern for C. He prefers $(1, 0)$ to $(\frac{1}{2}, 0)$ to $(0, 0)$. The problem is that it is only the end-state, C's welfare, that counts: his own part in providing for that welfare is recorded as a loss. He has none of what Margolis (1982) has called "participation altruism": he derives no satisfaction from the act of contributing itself.

If A and B are both calculating altruists and if they have to decide on their contributions independently of one another, then we may fall either in a Prisoner's Dilemma or in a game of Chicken. As the example has been set up, it is a case of Chicken. A would prefer B to meet C's needs; but if he really believes that B is going to pass by, then he will meet them himself: $(0, 1) > (1, 0) > (0, 0)$. For B, $(1, 0) > (0, 1) > (0, 0)$. There is no stable outcome to the game; each player makes a guess about the other's behaviour and acts accordingly.

To illustrate a PD, suppose instead that both A and B have only $\frac{1}{2}$ unit at their disposal and that they are both weakly al-

truistic. Both prefer $(\frac{1}{2}, \frac{1}{2})$ to $(0, 0)$ but for A $(0, 0) > (\frac{1}{2}, 0)$ and $(0, \frac{1}{2}) > (\frac{1}{2}, \frac{1}{2})$, whereas for B $(0, 0) > (0, \frac{1}{2})$ and $(\frac{1}{2}, 0) > (\frac{1}{2}, \frac{1}{2})$. Both then have an incentive to contribute 0 whatever they expect the other to do and we have a standard PD where the equilibrium outcome (0, 0) is suboptimal. In the two-person case, the psychology required to generate a PD seems unlikely to occur, since it requires that both A and B are willing to contribute $\frac{1}{2}$ if this has the joint result that C's need is completely met; on the other hand, neither is willing individually to raise C from 0 to $\frac{1}{2}$ or from $\frac{1}{2}$ to 1. However, the likelihood of a PD occurring rises sharply as the number of potential donors increases.

Thus a population of calculating altruists are liable to find themselves embroiled either in a game of Chicken or in a PD when faced with a group of needy people. The game will be Chicken if each of the altruists would in the last resort be willing to provide for the needs out of his own pocket; PD if he would only be willing to provide for some fraction of the needs as part of a joint endeavour. In either case, altruists of this kind ought to be willing to enter an enforceable agreement to donate. If the game is a PD, each can foresee that a suboptimal outcome (nobody donates) will arise. (I omit here discussion of voluntary co-operation as a consequence of repetition of the game. If the game is Chicken, there is a fair chance either of under-provision (nobody donates) or of inefficient over-provision).

We arrive in this way at a rationale for a welfare state as a means of extracting calculating altruists from their predicament. Voluntary donations may fail because, although each potential donor values the collective outcome of giving, none values it enough to donate of his own accord (each one values his own contribution negatively as a loss of resources that would be available for private ends). A more obvious way out might seem to reside in a collective contract where each person agrees to provide x on condition that a specific number of others do likewise. One might imagine specific charities to ask for conditional pled-

ges which would only be activated once the requisite number of names has been signed up. (This is the possibility envisaged by Nozick, though in relation to a population of reciprocal altruists). However, the solution is not theoretically viable: for the same reason that calculating altruists will be unwilling to donate independently, they may be unwilling to sign their conditional contracts: they may hope that the charity can find enough other donors. In conclusion, a population of calculating altruists should welcome a forcible system of transfers to the needy to which no one is exempt from contributing.

What about if each person has altruistic feelings, but there are differences in the direction of altruism - some are needs altruists; other are preference altruists? Here there are no free-rider problems; though there may be difficulties of coordination: few people are likely to want all the available resources spent in a single direction. But this is a different problem.

Calculating altruism may be challenged as a reasonable way of representing people's altruistic concerns. As Margolis (1982) and Sugden (1982) have pointed out, it would exclude commonly observed phenomena (charities that are supported by a large number of donors) and predict others that seem distinctly unlikely. There are large areas of altruistic behaviour which the hypothesis of calculating altruism cannot explain. So let's consider other forms of altruism attaching intrinsic value in some way to the act of giving: reciprocal altruism.

The reciprocal altruist is someone who is prepared to contribute to the welfare of the needy, but only on condition that the other members of a designated group also contribute: A will give to C provided B does also. Whereas for a calculating A the best outcome is (0, 1), for a reciprocal A the optimum is $(\frac{1}{2}, \frac{1}{2})$. Thus:

$$(\tfrac{1}{2}, \tfrac{1}{2}) > (0, 1)$$

$$(\tfrac{1}{2}, \tfrac{1}{2}) > (0, 0) > (\tfrac{1}{2}, 0) > (1, 0).$$

If B contributes $\frac{1}{2}$, A would prefer to reciprocate by giving $\frac{1}{2}$ himself rather then allow B to increase his contribution to 1. On the other hand, A will not contribute $\frac{1}{2}$ himself, much less 1, if B holds back. Thus (0, 0) is a possible outcome if each expects the other not to contribute.

The reciprocal altruist is clearly moved by a notion of fairness. Its presence might be accounted for in evolutionary terms, borrowing the idea that reciprocal altruism is a stable phenomenon, whereas loftier sorts of altruists are prone to exploitation by egoists and therefore liable to disappear in a competitive struggle for survival (see Dawkins (1976); McLean (1981)).

If A and B are both reciprocal altruists, they find themselves playing on Assurance game: Both prefer $(\frac{1}{2}, \frac{1}{2})$ to (0, 0). But each will contribute $\frac{1}{2}$ only on condition that he expects the other to reciprocate. Thus if they have to declare their contributions independently, the outcome depends on their mutual expectation. If they declare in sequence, with A going first, then B will play $\frac{1}{2}$ if A plays $\frac{1}{2}$ and 0 if A plays 0. What will A do? If he knows that B is also a reciprocal altruist then he will play $\frac{1}{2}$ and all is well. If he is uncertain about B's intentions, then his choice will depend on his estimate of the probability of B's contributing (p) and his relative valuation of the two payoffs. He will contribute if $p(\frac{1}{2}, \frac{1}{2}) > (1-p)(\frac{1}{2}, 0)$.

Therefore, a population of reciprocal altruists can arrive at an optimum by voluntary means provided they trust one another and can co-ordinate their behaviour. A compulsory welfare state is not necessary provided people are able to verify that others are pulling their weight. This suggests that compulsion might be needed to start the scheme up. This conclusion is vulnerable, however, to complications of at least two sorts. One is simply the presence of a small number of calculating egoists who will sabotage a voluntary scheme. (Indeed, calculating altruists will consider what chance their own failure to contribute would have on

the viability of the scheme as a whole). A second complication arises if the group in question is composed of people who are altruistic to different degrees. Each is willing to contribute 1/n of some ammont X if the others do, but X varies from person to person. Under these circumstances it will prove to be impossible to obtain voluntary contributions in excess of X_1/n, where X_1 is the value of X for the least altruistic member. (This is on the assumption that each contributor demands universal compliance with the level of donation he makes himself). Everyone else is deterred from supplying more of the altruistic good that they value by the reluctance of this person. A differentiated scheme of contributions would be more efficient. In short, relying on voluntary reciprocal altruism in a heterogeneous population leads to under-provision of need-satisfying goods. (Sugden, 1984).

In conclusion, consideration both of fairness and of efficiency will point reciprocal altruists towards compulsory financing of the welfare state, even though do not of themselves show that such compulsion would be legitimate.

There may be some doubts as to whether reciprocal altruism really counts as altruism. Isn't the real altruist the person who begins by giving $\frac{1}{2}$, waits for B to reciprocate, but in the last resort gives another $\frac{1}{2}$ if B fails to donate? (i.e. $(\frac{1}{2}, \frac{1}{2}) > (1, 0) > (\frac{1}{2}, 0) > (0, 0)$). These doubts spring from the "sympathy" interpretation of altruism. However, it is equally possible for altruism to take the form of acknowledged obligation: no one is obliged to take another's share of the burden upon himself. If A contributes while B does not, A is disadvantaged.

I move to a third character of altruist: the Kantian altruist (K.A.)(Collard 1978) who acts on a maxim which, if followed by everyone, would produce the outcome that he altruistically desires. He does so regardless of how other people are expected to behave. The Kantian altruist presumably prefers others to act on the maxim as well, but this has no effect on his own behaviour. So for a K.A.:

$(\frac{1}{2}, 0) > (0.0); (\frac{1}{2}, \frac{1}{2}) > (0, 1); (\frac{1}{2}, 0) > (1, 0).$

The last condition differentiates the K.A. from a fourth type, the Superkantian altruist for whom:

$$(\frac{1}{2}, \frac{1}{2}) > (0, 1); (1, 0) > (\frac{1}{2}, 0) > (0, 0).$$

The Superkantian not only does his own duty, but is prepared to do B's as well if B fails to contribute. He still, however, prefers $(\frac{1}{2}, \frac{1}{2})$ to $(1, 0)$. A person of whom the latter was not true would be a person whose "altruism" stemmed ultimately not from concern for the interests of C but from status-seeking or some such motive.

Clearly, a homogeneous population of Kantian altruists faces only co-ordination problems of the informational sort; there are no game-theoretical problems. Super-kantian altruists face only benign behavioural problems of the "after you" variety (i.e. if neither knows what the other's preference ordering is, they may be uncertain whether to donate $\frac{1}{2}$ or 1 and end up by over-supplying C).

It follows that Kantian altruists would have no need of a welfare state since they would be able to achieve the same result by a system of voluntary transfers. Note that K.A. are liable to be exploited by calculating altruists (See Lindbeck and Weibull (1988) and their discussion of the Samaritan's dilemma); so the former might welcome a welfare state as a means of compelling a recalcitrant minority of calculating altruists (assuming, that is, that they have $(\frac{1}{2}, \frac{1}{2}) > (\frac{1}{2}, 0)$).

This conclusion breaks down, however, if the Kantian altruists are needs altruists with differing interpretations of need. (Braybrooke, 1987). Each will then want to contribute to need as he identifies it, whereas a uniform, compulsory system will oblige him to satisfy a schedule of needs predetermined in

some way. However, this very possibility raises some doubts about the cogency of K.A. as a way of representing attitudes to the needy. The K.A., to recall, acts on a maxim that, if followed by everyone, would bring about the outcome that he altruistically desires, irrespective of his beliefs about how many others will actually do likewise. How intelligible a wiev is that? We need to investigate whether individual acts are likely to interrelate in such a way that the value of each depends on the character of the remainder. On the whole, welfare contributions are such that the interrelation is relatively significant. A single contribution, spread across a large number of people in need, will make very little impact. "Doing your bit" makes sense only if enough others are also doing theirs. (This is still a different attitude from that of the reciprocal altruist. The latter objects to contributing when others don't, seeing that possibility as unfair or exploitative. The attitude I am now describing is one of wanting to do "the right thing" regardless of others, but understanding that what "the right thing" is may depend on how others behave. None the less, the practical effects may be rather similar).

Kantian altruism makes sense where, by acting in a certain way, I can confer a visible good or avoid a visible harm. Deprived of this certainty, people who would be kantian are likely to behave in ways that suggest calculating or reciprocal altruism. The problem is not necessarily that people don't have the right moral capacities, but rather that the way in which they see their relationship to others (both to other donors and to the needy) doesn't bring those capacities into play.

In conclusion, Kantian altruism could only work effectively in a small and homogenous group of people. In a recent paper, Bordignon (1990) has shown that private provision of a public good is in general inefficient even if individuals follow such a strong ethical rule as Kantian behaviour. Useless to say, this is not a sufficient condition for public provision unless it is shown that governments can do better.

7. Teleological and deontological approaches to morality

Looking back at the history of the controversy over the relation between the rational pursuit of interest and moral rules, it is no wonder that the problem of the justification of moral rules has seemed so intractable. If rational interest requires one to violate moral side constraints when this can be done with impunity, then it is understandable that theorists have been driven to seek some other way to justify such rules.

The history of this justificatory problem is, as a matter of fact, a series of dialectical encounters between those who, like Hobbes and Hume, attempt to make the notion of rational interest central and those who, like Kant, sensing the inadequacy of any such account, are driven to postulate that moral rules must be grounded in some radically different fashion. For the former group, the organizing presupposition is that rationality provides an unproblematic basis for the justification of moral rules; but the residual problem they are unable to resolve is simply that this form of justification, as it has traditionally been construed, cannot be stretched far enough. For the latter group, the organizing presupposition is that moral rules must be shown to be more than simply conditionally binding; but the residual problem that they are unable to resolve is simply that nothing they have tried to offer in place of a theory of rational interested choice seems very compelling. At root, these two traditions face a common problem, for both have saddled themselves with one and the same limited restricted theory of rationality. An alternative theory of extended rationality, might, one would hope, provide a way to bridge the gap between moral choice and rational interested choice.

There is, however, a fundamental difficulty in the way towards the construction of such a bridge. It is best described with reference to a problem - perhaps the crucial problem - in the Rawlsian theory of justice.

The moral virtues - such as justice, altruism and so on - as seen by Aristotle derive their meaning from the teleological framework of thought that links them to the good, at least as understood by human beings. Kant reversed the order of priority, attributing more importance to what is right and less to what is good, so that virtues such as justice or altruism acquire their meaning from a deontological framework of thought.

Rawls clearly belongs to the tradition of Kant rather than that of Aristotle. However, whereas Kant's conception of justice applied primarily to the relations between individuals, Rawls' applies primarily to institutions. This deontological approach to morality has managed to maintain itself at the institutional level only by basing itself on the fiction of a social contract. This conjunction of a resolutely deontological approach to moral questions and the contractarian approach at institutional level is the central subject tacked by Rawls.

Is the connection contingent? Does a deontological approach to morality logically form an integral part of a contractarian procedure when virtues are applied to institutions rather than to individuals - as the virtue of justice is? What sort of link exists between a deontological viewpoint and a contrarian procedure?.

I believe that this link is not at all contingent, in that the purpose and function of a contractarian procedure is to ensure the primacy of the right over the good by substituting the deliberative procedure itself for a commitment to a supposed common good. According to this thesis it is the contractual procedure that is assumed to engender the principles of justice. If that is indeed the main point, then the entire problem of providing a satisfactory explanation for the idea of justice hinges on the difficulty of determining whether a contractarian theory can replace, with a procedural approach, any attempt to base justice on a few prior convinctions concerning the common good of the politeia.

It is to this central question that Rawls provides the best answer of recent times. He attempts to solve the problem left unsolved by Kant: how to proceed fron the first principles of morality - autonomy in its etymological sense (liberty, being rational, lays down the law for itself as a rule for the universalization of its own maxims for conduct) - to the social control by means of which a large number of people surrender their external liberty in order to recover it as members of a republic? In other words, what is the link between autonomy and the social contract? Such a link is assumed not proved by Kant.

Now, if it were possible for Rawls' endeavour to succeed we would have to say that a pure procedural conception of justice can have a sense without any assumption about the good and that it can even liberate the right from the tutelage of the good, firstly as it relates to institutions and secondly, by implication, as it relates to individuals.

My main objection is that a moral sense of justice, based on the golden rule ("Whatsoever ye would that men should do to you, do ye even so to them") is always assumed by the pure procedural proof of the principle of justice. This objection does not amount to a refutation of Rawls' theory of justice. On the contrary, it comes down to a sort of indirect defence of the primacy of this moral sense of justice in that Rawls' construct borrows its underlying thrust from the very principle he claims to generate by his pure contractual procedure. In other words, the circularity of Rawls' argument actually constitutes, in my view, an indirect plea for the pursuit of an ethical foundation for the concept of justice: a purely procedural conception of justice cannot replace the ethical foundation of our socio-political sense of justice.

The entire book by Rawls is an attempt to shift the emphasis from the question of foundation to a question of mutual agreement, which is the very essence of any contractarian theory of justice. Rawls' theory is undoubtetly a deontological theory in that it is opposed to the teleological approach of utilitarism

(he is not thinking of Plato or Aristotle), but his brand of deontology has no transcendental foundation (as in Kant). That is because the purpose of the social contract is to derive the content of the principles of justice from a fair procedure without making any commitment to a few objective criteria of what is just - for fear of ultimately reintroducing, according to Rawls, some assumptions with regard to the good. Thus the entire book seeks to provide a contractarian version of Kantian autonomy. The main purpose is to replace a foundational approach to the question of justice as far as possible by a procedural approach. Hence the constructivist form the book shares with the rest of the contractarian tradition. When the just is subordinate to the good it must be discovered; when it is engendered by procedural means it must be constructed: it is not anticipated but is assumed to follow from deliberation in conditions of complete fairness.

Now, it is my contention that a procedural conception of justice provides at best a rationalization of a sense of justice that remains a presupposition. The millstone of any contractarian theory may be that it derives from a procedure approved by all the very principles of justice that, paradoxically, have already motivated the search for an independent argument.

This ambiguity ultimately concerns the role of rational arguments in ethics. Can they be made to replace prior convictions by devising a hypothetical deliberative situation? Or is their role rather to throw a critical light on prior convictions?

It seems to me that Rawls is trying to have the best of two worlds: to construct a pure procedural conception of justice without losing the security provided by a "reflective equilibrium" between conviction and theory. In my view it is our intuitive understanding of the just and the unjust that ensures the deontological aim of the allegedly independent argument (including the maximin rule). Once removed from the context of the golden rule the maximin rule would remain a purely prudential argument typical of any bargaining process. Neither the deontological

intention nor even the historical dimension of the sense of justice is simply intuitive: they are the result of a long process of Bildung originating in the Graeco-Roman traditions. Divorced from that cultural background the maximin rule would lose its ethical character.

But this first suggestion concerning the epistemological status of rational arguments in ethics has no sense except when taken in conjunction with the second. We cannot do without a critical assessment of our supposed sense of justice. The task would be to discover which components or aspects of our considered convinctions require the constant eradication of prejudices and ideological biases. One may ask whether it is not pure utopia to trust in the rationality of ordinary citizens, that is to say their capacity to put themselves in the place of others or better still to trascend their place. But without such an act of trust the philosophical fable of the original position would be no more than an unbelievable and irrelevant hypothesis.

8. Do views about human nature matter?

Frank (1988) has shown that the modern presumption of a severe penalty for behaving morally is utterly without foundation. The idea is not that self-interest is an unimportant human motive, but that material forces allow room for more noble motive as well. We have always known that society as a whole is better off when people respect the legitimate interests of others. What has not been clear is that moral behaviour often confers material benefits on the very individuals who practice it. That such benefits exists is an encouraging piece of news. Largely as a result of the self-interest model's influence, our bonds of trust have taken a heavy beating in recent years.

More important, our beliefs about human nature help shape human nature itself. What we think about ourselves and our possi-

bilities determine what we aspire to become. Here the pernicious effects of the self-interest theory have been most disturbing. It tells us that to behave morally is to invite others to take advantage of us. By encouraging us to expect the worst in others, it brings out the worst in us. On the contrary, moral predispositions can be advantageous and above all they have beneficial effects on our behaviour. However, in order for such predispositions to be advantageous, others must be able to discern that we have them. This is a point which is seldom made.

Some may object that the prospect of material gain is not a proper motive for adopting moral values. This objection, however, miss the point. Satisfaction from doing the right thing must not be premised on the fact that material gains may later follow; rather, satisfaction must be intrinsic to the act itself. Otherwise a person will lack the necessary motivation to make self-sacrificing choices; and once others sense that, material gains will not, in fact, follow. Under the commitment assumption, moral values do not lead to material advantage unless they are heartfelt. Moreover, if it is true that unopportunistic behaviour is often beneficial, it is surely useful for people to know this. This helps to counteract the opposing tendencies encouraged by the selfish model.

References

Akerlof G., 1980, "A Theory of Social Customs of Which Unemployment May Be One Consequence", The Quarterly Journal of Economics.

Becker G.S., 1981, A Treatise on the Family, Cambridge: Harvard University Press.

Bernheim D., Stark O., 1988, "Altruism within the family reconsidered: do nice guys finish last?", The American Economic Review, 78, 1034-45.

Bordignon M., 1990, "Was Kant right? Voluntary provision of public goods under the principle of unconditional commitment", Economic Notes, 3, 342-72.

Braybrooke D., 1978, Meeting Needs, Princeton: Princeton University Press.

Broome J., 1978, "Rational Choice and value in economics", Oxford Economic Papers, 30, 313-33.

Collard D.A., 1978, Altruism and Economy, Oxford: Martin Robertson.

Coleman J.S., 1983, "Free Riders and Zealots", in Sodeur W.(ed.), Okonomische Erklarungen Sozialen Verhaltens, Bad Homburg, Verlag Schriften.

Dawkins R., 1976, The Selfish Gene, Oxford: Clarendon Press.

Elster J. (Ed.), 1986, The Multiple Self, Cambridge: Cambridge University Press.

Elster J., 1984, Ulysses and the Sirens (rev. edn.), Cambridge: Cambridge University Press.

Frank R., 1988, Passions within reasons, New York: Norton & Co.

Gauthier D., 1986, Morals by Agreement, Oxford: Clarendon Press.

Green H., 1971, Consumer Theory, Harmondsworth: Penguin.

Hammond P. (1982), "Utilitarism, uncertainty and information", in Sen A. and Williams B. (1982), 85-102.

Harsanyi J.C. (1982), "Morality and the theory of rational behaviour", in Sen A. and Williams B. (1982), 39-62.

Hargreaves Heap S., 1989, Rationality in Economics, Oxford: Blackwell.

Hirschman A.O., 1984, "Against Parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse", Bulletin: The American Academy of Arts and Sciences, vol. 37, 11-28.

Laffont J.J., 1975, "Macroeconomic constraints, economic efficiency and ethics: an introduction to Kantian economics", Economica.

Lindbeck A., Weibull J.W., 1988, "Altruism and time consistency: the economics of fait accomply", Journal of Political Economy, 96, 1165-1182.

Margolis H., 1982, Selfishness, Altruism and Rationality, Cambridge: Cambridge University Press.

Mc Clennen E.F., 1990, Rationality and Dynamic Choice, Cambridge: Cambridge University Press.

McLean I., 1981, "The Social Contract in Leviathan and the Prisoner's Dilemma Supergame", Political Studies, 29, 339-51.

Miller D., 1989, Market, State and Community, Oxford: Clarendon Press.

Parfit D., 1984, Reasons and Persons, Oxford: Clarendon Press.

Rawls J., 1971, A Theory of Justice, Cambridge: Harvard University Press.

Sen A., 1967, "Isolation, assurance and the social rate of discount", Quarterly Journal of Economics, 81, 112-125.

Sen A., 1977, "Rational Fools", Philosophy and Public Affairs, 6, 317-344.

Sen A., 1987, On Ethics and Economics, Oxford: Blackwell.

Sen A., Williams B. (eds.), Utilitarism and beyond, Cambridge: Cambridge University Press.

Steedman I. and Krause U., 1986, "Goethe's Faust, Arrow's Possibility Theorem and the individual decision-taker", in Elster J. (1986).

Stigler G.J., 1982, "The ethics of competition: the friendly economists", The Economist as a Preacher and Other Essays, Chicago, University of Chicago Press.

Sudgen R., 1982, "On the Economics of Philantropy", Economic Journal, 92, 341-50.

Sudgen R., 1984, "Reciprocity: the supply of public goods through voluntary contribution", Economic Journal, 94, 772-87.