Social relations instead of altruistic punishment. Comments on Ernst Fehr's altruism research

Leist, Anton

Abstract: Experimental economists have been trying for some time to discover the laws of behaviour in micro-social situations. Fehr's experimental research on altruistic behaviour attempts to correct the egoistic version of the concept of homo oeconomicus by resorting to the notion of altruistic dispositions. This article discusses Fehr's results from two points of view, namely in regard to the conception of social acting that is associated with altruism, and in regard to the research strategy associated with the laboratory method. The author argues that Fehr's concept of altruism distorts the representation of social acting and that, due to a lack of clarity concerning the motives of action, Fehr's empirical results pertain to phenomena of social recognition rather than to altruism. The charge against the research strategy is that it makes visible only local phenomena within the far wider field of general social conditions. Therefore, this approach presupposes more than it can explain.

Analyse & Kritik 27/2005 (© Lucius & Lucius, Stuttgart) p. xxx–xxx

Anton Leist

Social Relations instead of Altruistic Punishment

Comments on Ernst Fehr's Altruism Research

Abstract: Ernst Fehr's experimental research on altruistic behaviour aims at superseding the classical *homo oeconomicus* in micro-economic behaviour theory. This essay discusses Fehr's results from two points of view: first, in regard to the understanding of social action associated with the term "altruism"; second, in regard to the 'anthropological' strategy of research that is based on the laboratory method. Against the emphasis on altruism it will be argued that it misleads into providing a distorted description of social acting, and that, due to insufficient clarity about motives for acting, Fehr's empirical results give evidence not of altruism but rather of phenomena of social recognition. The objection against the anthropological strategy will be that it makes visible only local phenomena within prevailing social conditions and that it thus assumes more than it explains.

1. Is it Humans who Are Altruistic?

Reading Fehr's and his colleagues' studies is both impressive and confusing. It is impressive, as these authors promise nothing less than empirically proving—not just postulating as many philosophers and quite a few sociologists have done—a model of behaviour that is different from and incompatible with the classical *homo oeconomicus*. It is confusing, as this alternative model of behaviour remains unclear and, in the end, shares important features with *homo oeconomicus*. Both attractiveness and confusion are implicated in the two most important complexes of statements for which Fehr and his colleagues strive with their model of behaviour, namely, altruism and its socially transcendent universal validity. Both of these theoretical goals are necessary for a 'model of behaviour', for without tendencies of behaviour to be depicted it would not be possible to characterize behaviour, and without presuming context-free ways of behaviour it would not be possible to say anything about the social relevance of acting within society, from whose existence and possible effect on behaviour the empirical analysis abstracts. For the sake of brevity, I will, in the following, call the first theoretical goal the *altruism thesis*, the second one the *anthropology thesis*. "Altruism thesis" refers to all of Fehr's and his colleagues' attempts to understand social relations on the basis of whatever is the contrary of egoism.[1]

[1] In the following, I will omit the expression "and colleagues" so that the personal reference to Fehr will always mean the whole group working at the Zurich 'Institute for Empirical Research in Economics'.

"Anthropology thesis" summarizes all attempts to discover a socially context-free law of acting given the above concept of altruism.

I think that the goals of research named by these two theses are problematic, in a way typical of economic behaviour theory. This theory is problematic owing to its intention to design an anthropology of human behaviour which is supposed to be valid independently of context. Fehr's studies belong to this tradition in so far as they attempt to re-formulate *homo oeconomicus* as *homo reciprocans* and thus promise a new, emended view of man's 'essence'. Using the label 'anthropological' perhaps produces a frown. Among empirical scientists, anthropology is considered to be speculation by philosophers. In part, the interest in socially context-free behavioural attitudes could also be marked by the notorious predicate "individualistic". But "individualism" would not cover the claim of investigating altruism especially in "humans" (Fehr/Fischbacher 2003) or in "human" co-operation (Fehr/Fischbacher 2004)—in contrast to investigating the altruistic tendencies of the behaviour of economics students in Zurich or Moscow (something Fehr has actually dealt with). The socio-biological approach on the basis of which Fehr develops his theory is doubtlessly tied to the encompassing anthropological assumption that socially relevant ways of human behaviour have developed in a socially transcendent way and thus in a way which is immanent to human nature.[2] Thus, 'altruistic anthropology' is a suitable if surprisingly antiquated-sounding title for Fehr's program. Humans, this is Fehr's welcome message, are a bit nicer than we have thought up to now. But altruism is a kind of human original power to which social structures should better be subjected instead of dominating human behaviour. Social and political institutions could help only in the *homo reciprocans'* coming out. Depending on what this reciprocating human really has to offer, this may not be a reason for apprehension. But still, this suggestion does share the modesty of every anthropological theory and in this regard it is not different from classical economic political theory. Among the politically relevant consequences of Fehr's model of behaviour will be the fact that abstract goals of providing direction as well as political ideals will become questionable, that social motivation by ideas and utopias will become doubtful, and that social cohesion will be reduced to local sanctions against selfish loners. The fact, however, that totalitarian societies were not only aware of local reciprocal altruism but were also easily able to enhance and exploit it should serve as a warning about the limitations of this theory.

2. Two Meanings of "Altruism"

Fehr's central expressions of "strong reciprocation" and "altruistic punishment" presuppose a certain way of understanding "altruism". In the economic literature, this word is often used as a vague collective term for all motivation and behaviour that is not 'selfish' or 'self-interested'. Thus, in most cases, neither

[2] "Why are humans so unusual among animals...? ...Human altruism goes far beyond that which has been observed in the animal world." (Fehr/Fischbacher 2003, 785)

the term nor the named behaviour are defined precisely but are only determined in relation to their opposites; frequently, it is uncertain how far their meaning is truly defined in this indirect way of merely relying on the contrast. Owing to these blurred conceptual relations, it remains largely unclear what altruistic acting means, and this observation also applies to Fehr's studies. Indeed, it is obviously not the case that conceptual clarity can be obtained by resorting to empirical data alone.

In order to understand what we mean by 'altruistic' acting, it will be necessary to first clarify an ambiguity that arises from the fact that the two expressions "selfish acting" and "self-interested acting" do not mean the same thing. It is important to render the term "altruistic" more precise at this point, as it has two meanings, one narrower, one wider, and the claims that are tied to them pertain to few or many kinds of human behaviour and, accordingly, they are weakly or strongly representative, respectively. Accordingly, the term is then compatible or incompatible with alternative descriptions of the structure of social acting.

A certain behaviour is 'selfish' if it serves only the actor in a morally doubtful way. Instead of 'morally doubtful' we could also say 'mentally unhealthy'.[3] 'Self-interested' behaviour, on the other hand, is deliberately understood as free of this connotation of immorality. Feeding oneself is self-interested but not selfish; reproducing is self-interested and usually beneficial to others, especially to the children and descendants. Correspondingly, there are two meanings of "altruism". Altruistic behaviour can be any behaviour that is not selfish, i.e., *also* self-interested behaviour. Or it can be any behaviour that is not selfish *and* not self-interested. Both in ordinary language and in the social sciences, altruism tends to be understood in this second sense, i.e., strictly, and is thus contrary to selfishness. While in the case of selfish acting one cares *only about oneself,* in the case of altruistic acting one *only cares about others* (with or without disadvantage to oneself). However, the biggest part of our acting is obviously neither selfish in the strict sense of the word nor is it strictly altruistic (in this sense) but is situated in the centre of self-interest. Thus, a wider understanding of altruism, in which interest in oneself and in others are joined, would be more appropriate. As I will suggest later, the interest in social recognition or in society or community corresponds to this intermediate meaning, in which self-interest and altruism are linked. But the common concept of altruism does not express this connection, as it has not become separated from the contrast to selfishness.[4]

---

[3] Social scientists are by their nature hesitant about including 'moral' or 'medical' terms in the definition of social attitudes. But simple behavioural relations or restrictions alone are not capable of pinning down egoism. Doing something 'for oneself' would not justify calling that behaviour egoistic. Saving one's life in a threatening situation is doing something 'for oneself', but surely not egoistic. Sharing goods with others in times of abundance is doing something 'for others', but surely not altruistic. So something normative has to be built into these terms.

[4] "... in most of my dealings with others of a cooperative kind, questions of benevolence or altruism simply do not arise, any more than questions of self-interest do. In my social life I cannot but be involved in reciprocal relationships, in which it may certainly be conceded that the price I have to pay for self-seeking behaviour is a loss of certain kinds of relationships. But if I want to lead a certain kind of life, with relationships of trust, friendship, and cooperation with others, then my wanting their good and my wanting my good are not two independent, discriminable desires." (MacIntyre 1967, 466) MacIntyre equates self-interest with selfishness,

Given this ambiguity of "altruistic", it is plain that the altruism thesis in Fehr's investigations leads to a kind of dilemma. Either using the term "altruism" is justified, when altruism in the stricter sense is used. But then it is unlikely that it labels a socially representative way of behaviour. Acting *only* for the sake of others is the exceptional case. Or "altruism" is taken in the wider sense, as including self-interest. In this case, the objection of lacking representation does cease to apply but so does the contrast, claimed by Fehr, to the standard theory of economics, which, plausibly, one should also express in terms of motivation by "self-interest" and not by "selfishness".[5] I will soon show that this latter possibility offers the better solution.

My way of labelling Fehr's social theory so far has been misleading in that it is not really the concept of "altruism" but that of "*strong reciprocity*" that is in the fore. Thus, in the next section, I will clarify this concept in relation to others (3). Following this, I will explain to what extent the investigated games leave the question unanswered as to which *motives* form the basis of the observed behaviour. I will try to make clear that the function of motives may not be neglected and that reference to motives is essential if one wishes to speak of altruistic behaviour. Beyond this, I will suggest that Fehr's data do not provide evidence of altruism, but rather of the phenomena of social recognition (4). Furthermore, the results do not necessarily prove anthropologically anchored social motives, but rather a strong dependence on social contexts, on social normalcy and on legitimate expectations. This context dependency undermines the anthropology thesis and as well the relevance of experiments that were carried out in social isolation (5).

Altogether, my remarks are meant as conceptually based observations by a non-empiricist, who instead of relying on his own empirical research can only refer to everyday experience of social behaviour. As, in the end, the acceptance of research carried out by the social sciences can originate *only from this background* of common experience, such a confrontation should prove to be illuminating for the more technical kind of research as well.

3. Strong Reciprocity—Altruism or Social Recognition?

In his various publications, Fehr labels the pattern of behaviour that he discovered in three different ways: "altruism", "strong reciprocity", and "fairness".[6] "Strong reciprocity" is an artificial expression without any clear meaning in

something I do not agree with. At least if the self is understood as socially structured, "self-interest" seems to me a term which can be used to point to a kind of motivation which is not as morally reprehensible as selfishness is.

[5] Fehr understands altruism as opposed to both 'selfish' and 'self-interested' ways of behaviour (see Fehr/Fischbacher 2002, C1; Fehr/Falk 2002, 691) but he often emphasizes the contrast to selfishness or—presumably the same—"extreme self-interest" (Fehr/Fischbacher 2002, C1). Sometimes there is also talk of "non-monetary" (Fehr/Falk 2002, 688) and "material" interests (Fehr/Fischbacher/Gächter 2002, 2), in contrast to which altruistic interests would then have to be "ideal". Unfortunately, these conceptual distinctions are not adhered to systematically.

[6] See especially Fehr/Fischbacher 2003 and Fehr/Rockenbach 2003 on "altruism",

ordinary language. As has already been explained, in everyday life we most commonly use "altruistic" for unconditional, voluntary, and one-sided giving, while with "fair" behaviour we denote, more or less, impartial or egalitarian behaviour. Someone judges in a fair way if she judges impartially, 'regardless of the person'; positively expressed, this means in an equal or egalitarian manner. Thus, altruistic and fair behaviour are *not only not identical*, they can also *not be subordinated* to one another, as altruism has nothing to do with equality and equality has nothing to do with voluntary giving. The term that most suitably covers the pattern of behaviour in question should therefore be "strong reciprocity". Now, what is this?

Fehr defines "*strong reciprocity*" as follows (Fehr/Fischbacher 2003, 785):

> "Strong reciprocity is a combination of altruistic rewarding, which is a predisposition to reward others for cooperative, norm-abiding behaviours, and altruistic punishment, which is a propensity to impose sanctions on others for norm violations."

The use of "altruistic" in the definiens is ambiguous. Does "altruistic rewarding" mean nothing else than the corresponding *behaviour*, or does it mean, in addition, a certain motivation, such as the readiness to help and to sacrifice something? In the first case, the predicate could be replaced by the description of the behaviour, in the second case, this would not be possible, since a special interpretation of the behaviour is involved. Two additional, lesser problems with the definition are the following. First, it does not include the kind of strong reciprocity that Fehr often reports: the advance on trust at the beginning of a co-operation.[7] Advance on trust does not mean rewarding co-operative behaviour and, without an advance of this kind, co-operation would not begin. Second, it should be noted that assuming the existence of norms may significantly restrict the meaningfulness of the behaviour, and that it contradicts the claim that norm-generating behaviour is being studied.[8] I will return to this point later (see 5). But on the whole, the question of how to deal with the ambiguity of behaviour with or without a description of motives in the case of altruistic behaviour seems to be the most important one.

In correspondence to this ambiguity, we may choose from two definitions of strong reciprocity, one *without* and one *with* altruistic motivation (which is called as such):

Fehr/Fischbacher/Gächter 2002 on "strong reciprocity" and Falk/Fehr/Fischbacher 2003 on "fairness". Often, these terms are combined with each other.

[7] E.g., the often repeated (in the face of widespread mass-unemployment strangely anachronistic) example of the *generously paying employer* who wishes to induce better work morale. See Fehr/Falk 202, 690; Fehr/Fischbacher 2002. In one passage, Fehr and his colleagues define strong reciprocity in a conditional way so that starting co-operation would *not* be an example of strong reciprocity: "The kindness of a strong reciprocator is thus *conditional* on the perceived kindness of the other player." (see Fehr/Fischbacher/Gächter 2002, 3, 4) According to this, the strongly reciprocal employer would not be able to pay in advance, as she is here not reciprocating.

[8] Sometimes Fehr aims at this claim: Fehr/Fischbacher 2004.

(SR1) Actors behave in a *strongly reciprocal* way toward each other if they voluntarily bear the cost of co-operation either for the benefit of the co-operation, or within its context, or in reaction to it.

(SR2) Actors behave in a *strongly reciprocal* way toward each other if they are morally (altruistically, fairly) *motivated* to bear the cost in favour of and within co-operation.

SR1 includes the taking of risks for the sake of future co-operation and accepting costs in the context of and particularly at the end of the co-operation as well as assuming costs following the co-operation, such as those of subsequently punishing the partner involved in the co-operation. How compelling it is to call these kinds of behaviour 'reciprocal' is a question which perhaps should not be dealt with more intensively. At least, all actors act by somehow 'referring' to others, even if they do not really expect the others to act toward them. Thus, the belatedly punishing actors do not expect that something happens to them, as they have given up their interest in co-operation and thus cannot be injured anymore.

Although the kinds of behaviour covered by SR1 are the same in regard to the readiness to bear costs, they could be united as *one* syndrome of behaviour *only* if the additional condition were satisfied that the actors are moved to bear the costs by a common or at least a coherent complex of motives. Otherwise, very different motives could be involved in producing a readiness to bear costs, a readiness that appears similar when viewed from the outside. Mostly in order to distinguish his explanations from more complex versions of self-interest, Fehr does indeed suggest that a common motivation lies behind these kinds of behaviour. Thus, we arrive at the second definition, SR2. However, SR2 leads to the problem of interpreting the observed kinds of behaviour, for which specific moral descriptions and terms can no longer be excluded, as they are essential to the meaning of behaviour in view of the actors' intentions. Just one example to indicate the relevance of the actors' intentions: altruism and fairness are—as I have already shown—not identical and can also not be subordinated to each other so that, in this way of expressing SR2, either a very special constellation of acting would be addressed or two different classes of motivation would have to be included. In any case, it is important for the actor's intentions how she understands a given concept, and her intentions or motives are important for the character of the action.

In how far does Fehr take the *problem of motivation* into account and thus transform evidence of SR1 into evidence of SR2? SR1 is open as far as motives are concerned and it does not include a clear description of altruistic acting. Indeed, Fehr's empirical evidence—as I will show in the following section—to a great extent supports *only* SR1 and thus leaves the question unanswered in how far the experimentally investigated behaviour is SR2-behaviour. The persistent claim in this research program that evidence of altruistic behaviour is being provided is therefore definitely misleading.

4. Which Motives behind which Actions?

SR1-acting is favourable in the sense that it results in increased co-operation. But this does not make it clearly altruistic yet, as accepting costs is compatible with both selfish and self-interested acting. If someone punishes someone else out of sadistic motives, she may both be willing to bear costs and, under special conditions, even support co-operation. This and other possibilities make it clear that motives are essential if one is to be justified in speaking of altruistic acting. *But almost all of Fehr's experiments are unclear in regard to his statements on motives.*

The most important motives behind actions which incur costs for the actor *and* which foster co-operation may be the following: actors act

(a) because they want to produce future advantage for themselves – **indirect altruism**;

(b) because they desire social recognition—**recognition**;

(c) because they want to foster the welfare of others/the public welfare—**direct altruism**;

(d) because they believe in fairness and equality—**fairness**.

According to Fehr, (self-interested) motivation (a) is out of question as the participants in one-shot ultimatum games meet only once. This does not exclude the possibility that some participants may think that they will meet again, but, presumably, this probability is very small. However, three different kinds of motivation remain, and only (c) deserves the predicate "altruistic". The question of how these three kinds of motivation are connected can be answered neither by a philosophical analysis of concepts nor by observing behaviour alone but only by a socio-psychological theory of moral motivation. In my opinion, the concept of altruism that Fehr uses in describing his investigations is not well suited for the development of clearly classifiable research hypotheses, as it too strongly adheres to the altruism/egoism-dualism, a dualism that inhibits conceptual differentiations. Even conceptually, moral motives form a far more complex field than a dualist pair or a simple continuum. The most important element of a socio-psychological theory will likely turn out to be social recognition (b), to which I will return soon.

In what way does Fehr prove that just the motives (c) and (d) are effective? Let us examine the three most elementary classes of situations that were investigated.

Ultimatum Game

The types of behaviour displayed in the ultimatum game provide evidence of the readiness to punish others even at a cost to oneself as well as of the ability to hypothetically anticipate this readiness or threat.[9] They offer evidence of the

[9] Variants of the ultimatum game are quoted in almost all of Fehr's publications and seem to be the empirical mainstay of his research program so far.

ability to correctly estimate and predict the fellow players' moral reactions. At least in the case of societies with standards of fairness, a consciousness of fairness is present, which enables the players to roughly estimate the threshold below which the fellow players will react punitively. This at least is the predominant description of what occurs in the ultimatum game. The respondent's readiness is called "altruistic" when she prevents payouts in cases in which the portion assigned to her drops below a certain threshold.

What can be said on the basis of this description about the motives for acting? *Almost nothing!* It is unclear, at least in the most simple version of the game, whether threats are implicated or whether just fair distribution is involved. In regard to the respondent's motives, a fair distribution would at best inform us that she does not consider fair distribution to be injurious. In cases, in which the responder punishes, it is left open by which of the three motives (b)–(d) she is driven. In my opinion, among the three motives, that of social recognition seems to be by far the most plausible one. The responder punishes because she feels scorned and injured. It is not clear why her punishment should be described as altruistic. In the case of single games, it is unclear what the benefit would be, for the sake of which the responder punishes. It is most uncertain that the punishment 'morally reforms' the distributor. This easily makes the acceptance of one's own loss for the sake of an uncertain advantage to others appear heroic. It seems implausible that the responder punishes only 'for the sake of fairness', i.e., that she punishes because unfair behaviour 'deserves punishment', regardless of the consequences. Thus, the *motives* for punishment 'for the sake of well-being' or 'for the sake of fairness' are excluded. And this makes it difficult to see what is supposed to make the responder's acting altruistic or fair.

If, therefore, the motives (c) and (d) are improbable, much speaks for the social recognition motive, for the wish to be recognized as an equal partner. This motive makes the readiness to punish in the case of refused acceptance directly understandable. However, the social acceptance motive is one of self-interest and not of altruism. To be socially recognized, to be accepted as equal, is deeply in our interest. Only if self-interest is narrowly and one-sidedly equated with monetary, material, selfish interests, can this be overlooked.

Public Goods

In contrast to the single ultimatum game, in public goods games, the players' behaviour is directed toward co-operative behaviour by "altruistic punishment" in the course of repeated rounds (see Fehr/Fischbacher 2003, 787; Fehr/Gächter 2002). The fact that strongly reciprocal players cannot count on advantages in future co-operation but that the results of learning from co-operation prove to be beneficial to others is considered evidence of altruistic behaviour.

> "The act of punishment does provide a material benefit for the future interaction partners of the punished subject but not for the punisher. Thus, the act of punishment, although costly for the punisher, provides a benefit to other members of the population by inducing potential non-cooperators to increase their investments. For this reason the act of punishment is an altruistic act." (Fehr/Gächter 2002, 139)

Concerning the motives, Fehr claimed that "strong negative feelings" were involved (Fehr/Gächter 2002, 139). Unclear as this description is, it does suggest that the punitive action was driven rather by motives of revenge or retaliation. An altruistic motive would have to be considered 'positive', not 'negative', i.e., positive in regard to general well-being. Thus, it again seems plausible that these 'negative feelings' express the feeling that one has been personally scorned, i.e., the social recognition motive again plays an essential role.

But would it not be possible in this case to *completely abandon* the question of *motivation* and to defend the use of "altruistic" on the grounds that it has been unambiguously proven that advantages for future partners will accrue, advantages that were caused by earlier punishment, even if they were unintended? The problem is that self-interested and even selfish behaviour *is* often beneficial to others. Present day investment in research by a pharmaceutical company will benefit future patients, but is not altruistic. The widespread purchasing of a certain good may eventually result in the decrease of the market price of this good so that future buyers will benefit from the increased purchasing. But still, we would not speak of altruistic acting on the part of the earlier buyers. Thus, the predicate "altruistic" cannot be meaningfully used without reference to motives.

Building up Trust

In one interesting class of games, two players co-operate in such a way that a 'trustee' may increase the financial holdings of an 'investor' and is then free to pay a certain share of the gain to the investor. Fehr has demonstrated that the amount of the payment by the trustee is correlated to the extent of the investor's plans for sanctions in the case of insufficient payment (in case of omitted payment these sanctions cannot nullify the extent of the gain): the payment morale rises if the investor voluntarily gives up the possibility of all sanctions. However, two thirds of the investors did not give up the possibility of sanctions even if they were informed about this effect (Fehr/Rockenbach 2003). This basically self-damaging behaviour on the part of the investors is considered as evidence of strong reciprocity.

This is an example of quite complex relations between trust and the threat of sanctions. However, the explanations offered (in Fehr/Rockenbach 2003, 139) seem to be contradictory. Why did the investors, who were informed of the negative consequences of holding on to the possibility of sanctions, nevertheless retain it? It seems irrational to argue that they would sacrifice their predictable loss for the sake of their wish to be able to inflict fair punishment (Fehr/Rockenbach 2003), since only by retaining the threat of sanctions do they make judgments on fairness necessary at all. Furthermore, the conflict between the co-operation-reducing effect of the sanctions and the same effect in the case of public goods is disturbing. According to Fehr, in the case of public goods, it is fairness that is supposed to explain this difference (Fehr/Rockenbach 2003, 140). But the same explanation would also have to be true of the relationships between two persons, unless it were invalidated there by personal feelings. If this were so, however, an explanation of the type (b), in which self-interest plays a larger role, would

have to be true also of the ultimatum game which is likewise a game involving two people. The total sum of these explanations does not seem to be coherent.

These examples provide evidence of a *considerable gap* between the empirical data on behaviour in the various games and the possible motives behind the behaviour. To me, it seems surprising that despite the great effort invested in the empirical research, the motives for action in the games remain largely in the dark. Thus, it is difficult to estimate the meaningfulness of these games in regard to moral behaviour in everyday life. In my opinion, the advantage that is claimed for laboratory experiments (see Fehr/Fischbacher 2003, 785) is strongly reduced by the lack of transferability to everyday life situations in society.

In addition, the fact that *special punishing* behaviour plays such an outstanding role in these games and that the peculiarity of this kind of behaviour is not methodically reflected, seems to be responsible for the considerable problem of interpretation. As a matter of fact, the punitive reaction is an implicitly ambiguous kind of action, and it is, therefore, meaningful only to a limited extent. It is ambiguous, as it cannot be unrestrictedly altruistic. The person who must be punished, must be injured in order to produce advantages to herself and to others. Thus, the motives for punishment must be complex and suggest ingredients such as revenge, retaliation, justice, but surely also self-interest. As far as the studies indicate positive, e.g., trust-building, social interests, they carry features of motivation such as (a) and (b). The claim of altruism in Fehr's investigations therefore rests, all in all, on a very narrow empirical basis.

However, motives such as revenge, retaliation, justice, and . . . punishment additionally point to a more socio-psychologically spacious and thus more appropriate context of description than to the narrow type of altruism, namely, to that of *social recognition*. Fehr overlooks this socio-psychologically far more important kind of relation, because in many of his experiments he conceptualises—in conflict with his commonly expressed criticism of 'material' self-interest—altruistic punishment only in terms of 'real money', while at least in wealthy societies, non-material sanctions, indeed, roughly speaking, the withdrawal of recognition and of acceptance, are more important than withholding money. The withholding of money is a special version of sanctioning behaviour, which among strangers is made possible by moral consciousness. "Social recognition" means accepting the other or being accepted by others as a partner. This acceptance is more elementary than a sense of fairness or equality in so far as it may but need not manifest itself as egalitarian acceptance. A feeling for the other for the other's sake does not need to be egalitarian, but may also be hierarchical. As an alternative to Fehr's terminology, one could thus speak of a *genuine social interest* in the form of social recognition.

In social psychology and in the philosophical theory of action, such a kind of interest is not unknown, and apart from the terms "recognition" and "acceptance", it can be described by terms such as "social identity", "social observance", "respect", "esteem", "social perception", and others.[10] What these

[10] In philosophy, the terminology of recognition in philosophy goes back to Mandeville, Rousseau and Hegel, in social psychology to Herbert G. Mead and Symbolic Interactionism: see for Hegel and Mead Honneth 1996. Whether there really is an 'economy of esteem', as

conceptual and theoretical attempts share, is that they point to a need to be taken socially seriously, a need that is—in contrast to the emotive play of giving in the case of altruism—crucial both for social persons and for relations and societies, because it is only on the basis of this need that social persons are constituted at all. Persons, like societies, are normatively structured, and this 'anthropological' insight cannot be captured by speaking of altruism.

5. Anthropology Thesis and Dependence on Norms

> "If we randomly pick two human strangers from a modern society and give them the chance to engage in repeated anonymous exchanges in a laboratory experiment, there is a high probability that reciprocally altruistic behaviour will emerge spontaneously." (Fehr/Fischbacher 2003, 785)

Along with the reliance on the method of investigating isolated and highly abstract game situations, statements such as the one just quoted suggest that an essential part of human behaviour is based on tendencies of behaviour that are deeply rooted in human nature. Undoubtedly, such tendencies really do exist. What, however, is of interest, is how particular or general they are and how much the social environment contributes to the fact that these tendencies express themselves in a certain fashion. In accordance with my previous criticism, it seems to me that humans rather possess a deeply rooted social interest, i.e., *the need of social relations* for their own sake, which is part of their self-interest and which is in conflict with egoism. *Identifying* this deeply rooted social motivation with 'altruism' and 'fairness' results in misleading or simply mistaken assertions.

If one were to claim that humans acted in an altruistic manner *independently* of the influence of social contexts, one would either mean altruism in the more restricted sense (giving), a type of altruism that is hardly meaningful for ordinary acting, or, misleadingly, non-selfish acting (altruism in the wider sense). Indeed, humans often do act 'socially', i.e., in the interest *also* of others; however, they do not consider such acting as altruistic but as socially interested. If one were to claim that fair acting was *independent* of the influence of social contexts, one would have to assume that moral equality was an anthropological tendency of behaviour and not a cultural achievement. The ethnological investigations of the ultimatum game involving 15 natural tribes disprove this assumption (see Henrich et al. (eds.) 2004). Egalitarian distribution in this game occurs more frequently in Western societies than in non-Western ones. This underscores the relevance of the distinction expressed earlier between game situations with and without assumed norms. The tendencies of behaviour observed by Fehr are not simply those of fairness but those of a *reaction against the violation* of the

---

claimed by Brennan/Pettit 2004, is an interesting question. It is not true that the acceptance of A must necessarily be withdrawn from some B, and vice versa. So the giving of esteem need not be linked to scarcity in the same way as money. Rather, esteem is inherently universalistic.

norms of fairness and may turn out very differently depending on the different understanding and validity of these norms.

However, abstract situations of unequal sharing are less relevant in (Western) societies than situations of *unequal acceptance* in certain situations in life, for personal identities, and for social roles. The fact that the ideas of sharing are nearly egalitarian in the abstract ultimatum game does not tell much about the normative relations between concrete humans, such as men and women, as most possibilities for acting and reacting are normatively structured and influenced according to social roles and social perceptions. Therefore, the question becomes acute what the results of these games, which were carried out with great expenditure, contribute toward an understanding of everyday acting.[11] Women with an egalitarian image of themselves will react punitively in response to non-egalitarian offers, if they can afford to do so socially. And women without an egalitarian image of themselves will react positively to the same offers. It would be interesting to find out if the reaction could be *independent* of the normative self-image, thus indicating an independent 'anthropological' tendency of behaviour. However, the examples just indicated tend to prove the opposite: for these women, the self-image also determines the reaction. In so far as these women adhere to different moralities in different societies, social structures do influence behaviour.[12]

This points to a general *competition of explanatory strategies*. One may attempt to provide evidence of anthropological *dispositions* by abstracting from particular prevailing conditions. If the players act independently of their social roles and of the context of their personal situations, they can only meet 'as humans' and thus show relatively general dispositions. Social determinants have been abstracted from. Or we could try to distinguish the more or less general *types of relations* and assume that they mostly determine motives and actions. Only in a second step would we examine how the types of relation themselves are broken, updated, varied, and designed. According to a common sociological assumption, all social action occurs against the background of well-functioning and normatively structured types of relations. In my opinion, it is not difficult to recognize the advantages of this second explanatory strategy over Fehr's method.

The abstraction method and the socio-biological strategy must contend with the difficulty that there is *hardly* a game constellation which *does not* influence motives. Even highly abstract and, when measured up against everyday life, artificial games send messages to the participants and thus influence their actions. The fact—quite conspicuous to non-economists—that in the ultimatum game 'real money' is used, points to a special kind of behaviour in which the intensive aversion against other than egalitarian distribution is not surprising. In contrast to Fehr's interpretations, I think that punitive reactions usually arise out of a combination of self-interest and a lust for revenge ('negative emotions'). Instead

[11] For this, I assume that it is not necessary to disprove the model of strictly selfish behaviour, which, presumably, is no longer widely shared even among economists.

[12] The interdependencies between personal and social identities, moral rules and roles are surely complex and cannot be reduced to one-sided relations. Recognition again points to a socio-psychological mechanism and to the need to investigate the extent of these dependencies. Talk of altruism is again useless for this purpose.

of proving altruistic tendencies, the reactions rather seem to be (in the sense of the second explanatory strategy) the result of insecure and ambivalent social relations. This assumption is supported by the observations on the failure in trust building (Fehr/Rockenbach 2003). The phenomenon of suppressing altruism by types of action that are connected to the commercialisation of goods and relations, have long been known (see Titmuss 1970) and are thus not surprising.

Instead of allowing only the single actors to form the mutual 'setting', as in the study of the effect of less strongly reciprocal actors on many egoists (Fehr/Fischbacher/Gächter 2002, 11–20), a *normative concept of setting* thus seems to be preferable, with the help of which one may distinguish between different action situations, tasks, goals, and conflicts, and particularly supposed norms. Non-economists are not astonished by the fact that people with normative convictions act and react accordingly. It would be far more interesting to know how certain kinds of relations are exactly structured and which expectations and types of behaviour build up and become manifest in the context of these relations. Fehr's studies are guided by the assumption that scientific explanations of behaviour consist of classifications of motives and actions under certain conditions. If, however, *the prevailing conditions*—beyond the partners' concrete actions—*influence* motives and actions to a great extent, then individualistic explanations of this kind are uninformative and should be complemented by '*structural explanations'*.

However, structural—in contrast to individualistic—explanations reject two ideas that are closely connected to social micro-theories such as Fehr's. First, the notion that there are general anthropological motives of behaviour which could be identified in changing contexts and which would constitute a recognizable profile of behaviour. Against this assumption, the structural explanation stresses the effects of the different types of relations. It seems to me that in regard to Fehr's investigations it may be claimed that the individualistic explanation is not superior to the structural one. Second, and more important, the explanation that resorts to types of relations corrects a persistent blindness of the individualistic explanations, according to which moral behaviour must necessarily be attributed to motives of 'for me' or 'for others' instead of being explained on the basis of the types of relations. Elevating the ultimatum game to the status of a paradigm of social relations displays features of such a misunderstanding of social relations. The part of our relations that is most important in our lives *is not of the kind* that involves sharing. A theory whose empirical base is sharing is thus in danger of misunderstanding important parts of the social and moral world.[13]

Bibliography

Brennan, G./P. Pettit (2004), *The Economy of Esteem*, Oxford

Falk, A./E. Fehr/U. Fischbacher (2003), On the Nature of Fair Behavior, in: *Economic Inquiry 41.1*, 20–26

[13] I would like to thank Michael Baurmann, Vilem Mudroch, Mark Peacock, and Michael Schefczyk for their helpful comments.

Fehr, E./A. Falk (2002), Psychological Foundations of Incentives, in: *European Economics Review 46*, 687–724
—/U. Fischbacher (2002), Why Social Preferences Matter. The Impact of Non-Selfish Motives on Competition, Cooperation and Incentives, in: *Economic Journal 112*, C1–C33
—/— (2003), The Nature of Human Altruism, in: *Nature 425*, 785–791
—/— (2004), Social Norms and Human Cooperation, in: *Trends in Cognitive Science 8.4*, 185–190
—/—/S. Gächter (2002), Strong Reciprocity, Human Cooperation, and the Enforcement of Social Norms, in: *Human Nature 13.1*, 1–25
—/S. Gächter (2002), Altruistic Punishment in Humans, in: *Nature 415*, 137–140
—/B. Rockenbach (2003), Detrimental Effects of Sanctions on Human Altruism, in: *Nature 422*, 137–140
Henrich, J. et al (eds.) (2004), *Foundations of Human Sociality. Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford
Honneth, A. (1996), *The Struggle for Recognition*, Cambridge/MA
MacIntyre, A. (1967), Egoism and Altruism, in: P.Edwards (ed.), *Encyclopedia of Philosophy*, Vol. 2, 462–466
Titmuss, R. M. (1970), *The Gift Relationship*, Harmondsworth