Analyse & Kritik 27/2005 (© Lucius & Lucius, Stuttgart) 188–196

Mark S. Peacock/Michael Schefczyk/Peter Schaber

Altruism and the Indispensability of Motives

Abstract: In this paper we examine Fehr's notions of "altruism", "strong reciprocity" and "altruistic punishment" and query his ascription of altruism. We suggest that, *pace* Fehr, altruism cannot be defined behaviourally because the definition of altruism must refer to the motives of actors. We also advert to certain inconsistencies in Fehr's usage of his terms and we question his explanation of altruism in terms of 'social preferences'.

0. Introduction

It is among the standard objections to utilitarian thought that it demands an unrealistic degree of altruism. An early utilitarian like John Stuart Mill was aware of this problem. He was concerned that utilitarianism made unreasonable demands on people's willingness to sacrifice their own happiness for the general interests of society. The sort of altruism Mill had in mind was of a more limited nature: in normal circumstances the moral man is asked to attend only to the happiness of the people close to him. Only those people whose decisions have a large impact on the whole society should consider public utility.

This idea of limited altruism is common in the philosophical literature. In normal circumstances, people ought to help others as a matter of morality; most moral philosophers also contend that altruism is a necessary ingredient of a good life; but they, like Mill, usually mean an altruism which extends to people one feels close to, the usual suspects being children, parents, friends, colleagues and neighbours (see, for instance, Mackie 1977, 167). If what Ernst Fehr and his colleagues[1] call "strong reciprocity" is, as they are wont to aver, a form of altruistic behaviour, then a conception of human motivation that is popular in moral philosophy must be revised; for strong reciprocity is a form of unselfish behaviour which benefits neither the acting person nor those close to her. If Fehr's empirical work has discovered altruistic behaviour in laboratory experiments, it is not of the limited kind that prevails in the literature on moral philosophy. Indeed, Fehr's empirical findings seem to support a form of unselfishness that is rather close to what utilitarianism and also, of course, some forms of Kantianism require: the willingness to do something for the public good at one's own cost. We are highly sympathetic to this outlook. However, we argue in this paper that strongly reciprocal behaviour is correctly described as a form of altruism only if it is motivated by appropriate concerns. We make our argument through a

[1] For convenience, we henceforth refer to Fehr and his co-authors using the collective term "Fehr". We in no way mean to belittle the contributions of Fehr's colleagues thereby.

conceptual analysis of Fehr's work, paying particular attention to his ascriptions of altruism to various forms of behaviour and inconsistencies therein.

1. The Indispensability of Motives

Fehr insists on a "behavioural—in contrast to a psychological—definition of altruism as being costly acts that confer benefits on other individuals" (Fehr/Fischbacher 2003, 785). By defining altruism by types of action, Fehr avoids mention of "psychological" entities, e.g., motives and intentions, which are empirically intractable and ambiguous. Nevertheless, he admits that altruism has something to do with motives. Indeed he states that "sound knowledge about the specific motives behind altruistic acts predominantly stems from laboratory experiments" (*ibid.*). One can, Fehr holds, infer from behaviour observed in the laboratory to the motives behind it.

Our central qualm with Fehr's definition is that it is a far cry from the common understanding of altruism; for it is part of the ordinary definition of altruistic behaviour that it be *motivated* by other-regarding concerns. Altruism is constituted by a regard for the interests of others and a motivational disposition to act in their interests. To ask what motives lie behind altruistic acts makes no sense because, in the ordinary understanding of altruism, the answer to the question is contained in the definition: an altruistic motive lies behind an altruistic act; if not, the act is not altruistic. Consequently, we must disagree with Fehr's contention that "[a]ltruistic behaviour in real-life circumstances can almost always be attributed to different motives" (*ibid.*). We would rather say: Real-life behaviour *which appears to be altruistic* can almost always be attributed to different motives; but if that behaviour is altruistic, it is so because the motive for the act is altruistic. This raises the question whether one can distinguish acts which merely appear to be, from acts which *really are*, altruistic. This leads, in turn, to the issue of scepticism: the difficulty in ascertaining whether an act which appears to be altruistic is actually motivated by altruistic concerns is the reason scepticism concerning altruism is common in moral philosophy, economic theory and everyday life. Fehr seems to circumnavigate such scepticism by defining altruism behaviourally rather than motivationally. But if he is to pursue his behavioural approach consistently, he would be better off abstaining from the attribution of motives; for motives resist complete characterisation in behavioural terms and so even an exact and complete description of an individual's behaviour does not allow one to draw unambiguous inferences to the motives of that individual.

An important advantage of the ordinary *vis-à-vis* Fehr's definition of altruism is that the former allows one to distinguish altruistic actions from other kinds of costly actions which give rise to positive externalities. To make this distinction, we must differentiate between the purpose of an action and its side-effect. We use the word "altruism" in order to characterise a person's purpose in acting and not the results of her act. Imagine, for instance, that a celebrity moves into your neighbourhood, thus raising the price of housing in the area. Let us assume that the celebrity wants to do his wife a favour but himself has liking for neither house nor neighbourhood. The celebrity's act *vis-à-vis* his wife qualifies as altruistic in our view and in Fehr's. According to Fehr's definition, however, the celebrity also acts altruistically towards his neighbours, for they too benefit from his costly act. We beg to differ with this assessment: the celebrity does not act altruistically towards the neighbours because he does not *intend* to benefit them; their benefit is an unintended effect of his purchase but not his motive for buying the house. If we are to characterise the act as altruistic with regard to the neighbours, we would have to appeal to 'psychological' information (intentions, motives): if the celebrity bears the cost of moving house *in order to* increase house prices of extant residents, then, but only then, could we make a case for calling his act altruistic. For Fehr, such information is irrelevant to the ascription of altruism. This becomes particularly clear in his 'biological' definition of altruism for which it is irrelevant whether "the act is motivated by the desire to confer benefits on others, because [biological] altruism is solely defined in terms of the consequences of behavior" (de Quervain et al. 2004, 1257). The biological definition seems to correspond to Fehr's behavioural definition (Fehr/Fischbacher 2003, 785). We agree that a behavioural/biological definition of altruism is apposite in biology; biologists study species which display altruistic behaviour which in all likelihood has no motives in the ordinary sense of the word. In contrast, humans do have motives. In personal as well as in professional contexts they invest a good deal of resources to find out what they are. The sceptical question whether someone did something (purely) for the sake of someone else or whether he pursued his self-interest in a camouflaged way is an ever-present question in human affairs. It has no counterpart in the world of non-human animals.

We conclude that a plausible definition should capture the intentional dimension of altruistic actions: X acts altruistically with regard to Y if and only if she bears costs of so acting *in order to* confer benefit on X. Let us turn now to "strong reciprocity".

2. Strong Reciprocity

The term "strong reciprocity" (henceforth SR) was coined by Gintis (2000) in the context of the research programme to which Fehr has contributed. SR has an advantage over "altruism": being a technical term, SR, unlike altruism, has no ordinary language counterpart with which its technical use can conflict. Fehr defines SR with what we call **DefSR**:

> "The essential feature of strong reciprocity is a willingness to sacrifice resources for rewarding fair and punishing unfair behavior *even if this is costly and provides neither present nor future material rewards for the reciprocator.*" (Fehr/Fischbacher/Gächter 2002, 3)

He distinguishes SR from both "reciprocal altruism" (Trivers 1971) and "altruism" (*sans phrase*) and elucidates the distinction with the example of a one-shot, sequential prisoners' dilemma in which the first-mover, A, decides whether to cooperate or defect, and B moves thereafter. A strong reciprocator "defects if A defected and cooperates if A cooperated" (Fehr/Fischbacher/Gächter 2002, 4). This is consistent with **DefSR** because B sacrifices resources in order to 'reward' A for good behaviour; B thereby attains a payoff of 5 (as does A in a cooperate-cooperate constellation) and forgoes a payoff of 7 (which she would have attained had she defected following A's cooperation).[2] Compare the reciprocal altruist: she, by definition, only cooperates if future returns are to be had thereby; hence she will defect in a one-shot, sequential prisoners' dilemma come what may. The altruist, by contrast, cooperates come what may because she is, by Fehr's definition, *unconditionally* kind (*ibid.*).

In what follows, we advert to a possible inconsistency between Fehr's definitions and offer an interpretation which irons out the inconsistency. To do so, we examine **DefSR** which we divide into three:

> (i) The essential feature of strong reciprocity is a willingness to sacrifice resources for rewarding fair and punishing unfair behavior (ii) *even if this is costly and* (iii) *provides neither present nor future material rewards for the reciprocator.*

Let us analyze **DefSR** step by step: (i) leaves unclear whether rewarding fair and punishing unfair behaviour necessarily involves the sacrifice of resources (or, synonymously, whether it is *costly*). It could be that such rewarding and punishing is *not* always or necessarily costly but that strong reciprocators are willing (in the case that it is costly) to sacrifice the relevant resources. Let us call this interpretation (i)'. Alternatively, (i) could imply another interpretation, (i)", that rewarding and punishing are always necessarily costly and that strong reciprocators are willing to bear those costs.

The second clause, (ii) gives us a clue to the interpretation of (i); for the "even if", in (ii), must presumably be understood in its usual inclusive "whether or not" sense. Thus (ii), which is a condition on (i), leaves the question whether rewarding and punishing are costly open: *whether or not* rewarding and punishing are costly, strong reciprocators are willing to reward and punish at a cost to themselves. Our reading of (ii) thus supports (i)' and not (i)".

The final clause (iii) raises the question whether the "even if" of (ii) is to be carried over to (iii). This is the natural syntactical reading of (iii) and thus renders it: "*even if this ... provides neither present nor future material rewards for the reciprocator*". Reading "even if", once again, as "whether or not", this implies that the ascription of SR to an action is independent of the material rewards accruing to the reciprocator; there might be such rewards, but the important thing is that the reciprocator be willing to bear the costs of rewarding and punishing whether or not she receives material benefit (or expects to do so).

[2] We have placed "reward" in scare quotes because "reward" contains an evaluation of B's behaviour (rewarding being an appropriate thing to do in response to worthy deeds). "Reciprocate" is not evaluative; furthermore, it leaves B's motives for reciprocating open, whereas "reward" contains motivational information. In the context of Fehr's behavioural approach "reciprocate" fits better; to speak of "reward" he would have to delve more deeply into B's motives.

It is important to be clear on this in light of another definition which Fehr gives, namely of "reciprocity" (as opposed to SR). In a definition we term **DefRec**, Fehr states that reciprocity

> "means that people are willing to reward friendly actions and to punish hostile actions *although the reward or punishment causes a net reduction in the material payoff of those who reward or punish"* (Camerer/Fehr 2004, 56).

One question and one remark. First, the question: is reciprocity synonymous with SR? Here and elsewhere (Fehr/Gächter 2000, 160), Fehr does without the adjective "strong" but offers a definition very close to that of SR. "Very close", we say, but not "identical", for, secondly (and here is the remark), note that the "even if" in **DefSR** has been replaced by "although" in **DefRec**. "Although", unlike "even if", is categorical: the last, italicised part of **DefRec** must therefore be read: *in spite of the fact that those who reward or punish become materially worse off for doing so.* **DefRec**, in contrast to **DefSR**, is unambiguous: **DefRec** leaves us in no doubt *that* rewarding and punishing is costly whereas **DefSR**(iii) leaves the issue open.[3] Is then a mere *willingness* to incur net costs of rewarding and punishing a necessary condition of SR (whether or not these costs arise)? Or must one *actually* incur such net costs if one is to count as a strong reciprocator? In Fehr's experiments, punishing and rewarding are (by experimental construction) invariably costly. But is this merely because the experiments are so constructed or does this reflect the definition of SR? Our interpretation is as follows: in his experiments, Fehr constructs the 'hardest case', namely where rewarding and punishing are costly; in the experiments set up by Fehr, then, the "although" of **DefRec** applies. In the world outside the laboratory, however, other factors play a role: players can, for instance, meet more than once and reputations gained through cooperation; here, the 'even if' of **DefSR** applies, for although players might not have to forego material gain when they reward and punish (because, for example, the reputation effects of doing so outweigh the costs), they would reward and punish 'even if' such material gain were lacking. That they *would* do so 'even if' such gain were lacking is, of course, something Fehr confirms by observing his subjects in the experimental 'although' world. An experimental approach allows one to disentangle the myriad factors of the real world and isolate others. Hence the difference between "even if" and "although".

Having distinguished SR from altruism in the way we have described above, Fehr, to our surprise, defines SR in terms of altruism. This leads us to a discussion of "altruistic punishment".

[3] Things don't become clearer in Fehr and Gächter (2000, 160) in which **DefRec** and **DefSR** are combined; as in **DefRec**, Fehr and Gächter define "reciprocity" (not SR) but use the "even if" from **DefSR** rather than the "although" from **DefRec**.

3. Altruistic Punishment

The connection between SR and altruism is stated in a different definition of SR, namely:

> "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." (Fehr/Fischbacher 2003, 785; see also de Quervain et al. 2004, 1254)

In this section we argue that punishing others can be altruistic only if it is caused by appropriate motives.

As far as we can see, the term "altruistic punishment" makes its first appearance in Fehr's work in the context of a public goods experiment in which subjects punish free-riders. Punishment represents a contribution to a second-order public good which is in as much need of explanation as contributions to first-order public goods (Fehr/Gächter 2002 137). Why do subjects punish others at a net cost to themselves? According to Fehr, "negative emotions" caused by free-riding motivate punishers (a hypothesis he confirms by eliciting responses from experimental subjects regarding their "anger and annoyance" level when presented with various scenarios in which free-riding occurs). These negative emotions are the "proximate source" of punishment. We agree that negative emotions can play an important causal role in the punishment of norm violations; however, we have reservations about calling actions which are thus motivated altruistic without further qualification.

It is important to trace the context of Fehr's argument, here. Fehr (i) defines altruistic punishment according to two necessary and sufficient criteria (to wit, that punishment be (a) costly and (b) of no potential gain to the punisher). In public goods experiments, he (ii) observes punishment behaviour which satisfy these two criteria (thus making the punishment involved, for Fehr, *ipso facto* altruistic). He (iii) provides evidence for the hypothesis that negative emotions (caused by free-riding) are the proximate source of the observed punishment. Our question is: do the 'negative emotions' in (iii) provide support for the claim (i) that the punishment be 'altruistic', or is this merely a matter of definitional fiat? Consider a person, A, angry at B's free-riding behaviour and wondering whether to punish B. A wishes to 'teach B a lesson' and knows that others are likely to profit from her punishment, yet she faces an incentive to let others bear the cost of punishing B. A's anger together with the facts (a) she punishes B and (b) others benefit from her doing so are not, we hold, sufficient to term A's punishment altruistic. To talk of *altruistic* punishment, here, we would have to add: (1) that A expects others to benefit from her punishing B; (2) that if she punishes B *in order to benefit those others*; and (3) that she would not punish B if (2) were not the case (that is, if she were prepared to punish B for reasons other than her intention to benefit others). Note that we have expunged all mention of 'negative emotions'; whether anger is the cause of A's punishing B is immaterial for the ascription of altruism. It is worthy of note that Fehr's interpretation of the motive for 'altruistic punishment' bears a striking resemblance to what John Stuart Mill (Mill 1863/1998, 5.22) called the *natural feeling of retaliation* (which is at the bottom of our *sentiment of justice*). Fehr writes: "Most people seem to feel bad if they observe that norm violations are not punished, and they seem to feel relief and satisfaction if justice is established. Many languages even have proverbs indicating such feelings, for example, 'Revenge is sweet'." (de Quervain et al. 2004, 1254) But there is nothing obviously altruistic about a desire for revenge. So if the natural feeling of retaliation is a good candidate for the explanation of punishment (and we think it is), it is nevertheless insufficient for the ascription of altruism to acts of punishment for norm violations. Parenthetically, one may note that at times Fehr himself makes no mention of *altruistic* punishment when he describes similar public goods experiments (e.g., Gintis et al. 2003, 159-62); the term "punishment" alone suffices; the reader does not get the impression that the word is in need of adjectival embellishment. We in no way deny that the punishment which in part constitutes SR could be a manifestation of altruism. But if we are to argue that it is, we require a different understanding of altruism to that which Fehr offers; we require information about the intentions of punishers.

4. Social Preferences

One explanation of "altruistic punishment", according to Fehr, is that people have "social preferences ... which take into account the payoffs and perhaps the intentions of others" (Fehr/Fischbacher/Gächter 2002, 17). These are, very broadly, preferences for 'fair' allocations and norm-abiding behaviour. "From a theoretical viewpoint such preferences are not fundamentally different from preferences for food, the present versus the future, how close one's house is to work, and so forth." (*ibid.*) These preferences, Fehr argues, explain the altruistic punishment behaviour which partly characterises SR. And there, to cite the bard, is the rub, for if, by performing acts of punishment, strong reciprocators are acting on their preferences, then they are doing that which economists have, for a long time, held all people to do all the time, namely, strive for utility. That the *content* of strong reciprocators' preferences be different from that of a selfish person is, from an action-theoretic standpoint, neither here nor there. And with this explanation of SR, Fehr comes close to forfeiting his alternative to the self-interest model of human behaviour as we now explain.

Consider the term "social preference" of which we distinguish two interpretations. According to interpretation (I), to say that a person acts on the basis of a social preference means that (i) a person derives utility from punishing norm violation (or from rewarding norm-abiding behaviour) and that (ii) deriving utility is the reason for punishing/rewarding. Condition (ii) takes us back to a hedonistic version of self-interest theory. Fehr's most recent research in the new field of neuro-economics points in this direction. We think, however, that Fehr must rebut interpretation (I) if strong reciprocity is to be more than simply a type of preference satisfying behaviour (albeit with 'social preferences'). In order to challenge the self-interest model, one has to take seriously the possibility that people can be motivated by other-regarding concerns. Consider interpretation (II). According to it, a person acts (punishes/rewards) on the basis of social preferences if (i) she intends to promote the utility of others and if (ii) the fact that her act promotes the utility of others is her reason for action. Interpretation (II) is compatible with the first condition of interpretation (I) but not with its second condition. In what follows, we expatiate on this point.

To approach a full definition of altruistic behaviour, we note, first, our disagreement with one line of scepticism towards altruism: in our opinion, that an actor derive utility from an action is *not*, *pace* the sceptic, sufficient to exclude the ascription of altruism. Altruistic acts are not necessarily self-sacrificing if the latter means that they be performed at a cost to oneself; what matters for an altruistic act is that it be performed for the benefit of others. If the actor derives benefit from the action, then, we are inclined to say, so much the better; but it does not forbid us from saying that the act was altruistic. Hence, the question *whether* the actor derives benefit from the act is irrelevant to the question whether her act is altruistic. The decisive question is: What motivated the actor to act? If the motivation was a desire to gain utility (as in interpretation (I), condition (ii) in the previous paragraph), then we incline to the sceptical conclusion that the act was not altruistic. But if the motivation was a desire to help others (or otherwise benefit them), we are willing to admit the ascription of altruism irrespective of the gains which the actor acquires by helping others.

In suggesting that motives play a definitional role in altruism, our definition corresponds to what Fehr calls the "psychological" definition of altruism which "requires that the act be driven by an altruistic motive that is not based on hedonic rewards" (de Quervain et al. 2004, 1257). By abjuring the psychological definition, Fehr excludes motives from the definition of altruism and it is here that we must part company with him. But if he wants successfully to parry the sceptic's argument (that strong reciprocators be merely utility-seeking wolves in sheep's clothing), he must define altruism independently of the benefits accruing to the actor. Otherwise economists, amongst others, will query all the excitement about the behaviour of Fehr's experimental subjects when that behaviour is compatible with preference-satisfying behaviour (however 'social' those preferences may be). With suitable reformulation of the definition of altruism, Fehr's experimental results could be re-examined according to the motives of strong reciprocators. We believe that his findings will be relevant to the phenomenon of altruism in a way which is not only of interest to economists but also to philosophers.[4]

Bibliography

Camerer, C./E. Fehr (2004), Measuring Social Norms and Preferences Using Experimental Games, in: J. Henrich/R. Boyd/S. Bowles/C. Camerer/E. Fehr/H. Gintis (eds.), *Foundations of Human Sociality*, Oxford, 55–95

[4] Thanks are due to Bernd Irlenbusch and Elke Renner for comments on an earlier version of this essay.

de Quervain, D./U. Fischbacher/V. Treyer/M. Schellhammer/U. Schnyder/A. Buck/E. Fehr (2004), The Neural Basis of Altruistic Punishment, in: *Science 305*, 1254–1258

Fehr, E./U. Fischbacher/S. Gächter (2002), Strong Reciprocity, Human Cooperation and the Enforcement of Social Norms, in: *Human Nature 13*, 1–15

— /S. Gächter (2000), Fairness and Retaliation, in: *Journal of Economic Perpectives 14*, 159–81

— / — (2002), Altruistic Punishment in Humans, in: *Nature 415*, 137–40

— /U. Fischbacher (2003), The Nature of Human Altruism, in: *Nature 425*, 785–91

Gintis, H. (2000), Strong Reciprocity and Human Sociality, in: *Journal of Theoretical Biology 206*, 169–79

— /S. Bowles/R. Boyd/E. Fehr (2003), Explaining Altruistic Behavior in Humans, in: *Evolution and Human Behavior 24*, 153–72

Mackie, J. L. (1977), *Ethics: Inventing Right and Wrong*, Harmondsworth

Mill, J. St. (1863/1998), *Utilitarianism*, ed. by R. Crisp, Oxford

Trivers, R. (1971), The Evolution of Reciprocal Altruism, in: *Quarterly Review of Biology 46*, 35–57