Analyse & Kritik 27/2005 (© Lucius & Lucius, Stuttgart) 172–187

Hans Bernhard Schmid

'Nostrism': Social Identities in Experimental Games*

Abstract: In this paper it is argued that *a)* altruism is an inadequate label for human cooperative behavior, and *b)* an adequate account of cooperation has to depart from the standard economic model of human behavior by taking note of the agents' capacity to *see themselves and act as team-members.* Contrary to what Fehr et al. seem to think, the main problem of the conceptual limitations of the standard model is not so much the assumption of *selfishness* but rather the *atomistic* conception of the individual. A much-neglected question of the theory of cooperation is how the agent's *social identity* is determined, i.e. how individuals come to think of themselves and act as members of a group. Considering as an example one of Fehr et al.'s *third party punishment* experiments, I shall argue that the agents' identities (and thus the result of the experiment) are strongly influenced by the way the experiment is presented to the participants, especially by the collectivity-related vocabulary used in the instructions.

1. Snake or Hat?

In a graphic image used in the introductory chapter to the *Foundations of Human Sociality,* Ernst Fehr and Colin F. Camerer compare the role and scope of experimental games in the study of human sociality to that of a first outline in the process of an artist's conception of a painting. Just like rough sketches, experimental games are, as Camerer and Fehr put it, "reductions of social phenomena to something extremely simple" (Camerer/Fehr 2004, 85). By abstracting from contingent details and reducing complex phenomena to their essentials, experimental games allow for "comparability across subject pools" (ibid., 84), a feature of which the volume is itself a most impressive display.

As far as the art of drawing is concerned, however, reduction is always a bit of an adventure. In his *Little Prince,* Antoine de Saint-Exupéry has the narrator learning this the hard way when the little boy draws his 'Drawing Number One'—a giant snake digesting an elephant—only to learn that the adults mistake the lumpy blob with two lines tapering off to both sides to be a rendering of a hat! The lesson is that for all their simplicity, rough outlines need more interpretative work from the side of the beholder than more detailed pictures. And the more reductive a rendering, the more easily it is misunderstood. The question is: could this also be true of game experimental 'sketches'? In the following, I shall discuss

* I am grateful to Michael Schefczyk, Fabienne Peter, Raimo Tuomela and to the members of his research group for critical comments on earlier versions of this paper.

the most famous of Fehr et al.'s game experiments, his *third party punishment-experiments* which are designed to demonstrate how agents provide a *second order public good* by punishing unilateral deviation from the cooperation norm even where these agents themselves are not directly affected by the outcome.

The structure of one version of this experiment (Fehr/Fischbacher 2004, 72ff.) is as follows. The test person A is endowed with a certain amount of money by the experimenters, and she is offered the opportunity to spend parts of her fund on inflicting financial loss on another, randomly selected individual B. A does not know B's identity, she is assured anonymity by the experimenters, and she knows that there will be no further interaction whatsoever between herself and B in the future course of the experiment. All that A knows about B is B's decision in a previous round of the experiment, where B was in an interaction with yet another randomly selected individual C. In this first round of the experiment, both B and C are given the options either to keep the money they were endowed with by the experimenters for themselves, or to transfer their fund to the other, who would then receive three times the transferred sum (transfers are tripled by the experimenters). B and C have to make their decision simultaneously and without any prior communication. After the second round (where A comes into play), the entire experiment is repeated several times over, with the test persons being regrouped for each new series, and each person being in the position of A in the second round playing the role of B in another group in the first round. No pairing, however, is repeated, so that the experiment consists of *a series of one-shot interactions*. The structure of the game is common knowledge among the participants.

Roughly stated, the results of the experiment are the following: A tends to spend a surprisingly substantial part of her endowment inflicting financial loss on B, especially if in the earlier round B did not transfer his money, while C transferred hers (A's tendency to inflict loss on B was low if either B decided to transfer, or if both B and C decided not to transfer). Anticipating A's negative reaction to unilateral non-transferral in the second round of the experiment, B was more prone to choose the 'transfer'-option in the first round than she or he was when the experiment was limited to the first round only (interaction between B and C with no intervention by A).

To use Fehr's and Camerer's metaphor, this is the outline. The decisive question now is: what does it show? What *sense* can and should be made of it? What is the deeper *meaning* of the behavioral pattern revealed by this experiment? In Fehr et al.'s view, this represents the provision of a *second order public good*. The label they use for A's behavior is "altruistic punishment", and they claim it to be 'pro-social' and 'norm-enforcing', and a paradigmatic case of 'strong reciprocity'. None of these labels, however, is as obvious as Fehr et al. think.

a) The label "reciprocity", as attached to the behavior at hand, seems rather strange, since neither of the pairings is repeated, and B is given no opportunity whatsoever to retaliate against A.

b) The term "punishment" is questionable, too. For obvious reasons, A's

intervention cannot be interpreted simply as 'punishment' for B's refusal to transfer her fund to C in the first round of the experiment, since B's decision not to transfer was only 'punished' if C decided to transfer. In other words, *B's* non-cooperative choice was 'punished' by A only if it constituted a case of *unilateral deviation* from the cooperation norm. It is obvious, however, that B had no control whatsoever over whether his non-cooperative decision constituted a case of unilateral non-cooperation, since this depended on C's decision, which was not known to B, let alone under her or his control. This makes the label "punishment" highly questionable. In ordinary language, 'punishment' is an imposition of a penalty on somebody *for some wrongdoing on his part.* Whatever counts as right or wrong in a given context, the term implies some *doing* on the part of the punished person for which she or he is punished. A's 'punishing' behavior, however, was *not triggered by B's choice* but rather by the *outcome of the interaction* between B and C.

c) The label "altruism", as applied to A's behavior, is no more self-evident than "reciprocity" and "punishment". It is true that the behavior does not conform to the selfishness-assumption of classical economic theory, according to which because of the costs to A, no infliction of loss on B should have occurred. Thus, the behavior in question does not appear to be egoistic in a narrow sense of the word. Yet this does not, in itself, make it a case of *altruism.* Most experimental economists seem to tend to portray deviations from the principle of narrow self-interest as being of a 'benign' or 'pro-social' kind. As the true heirs of the enlightenment movement and its positive picture of human nature, they have often found humans to depart from their self-interest out of *altruism*, *inequality-aversion*, or generally *fairness-orientedness* (most obviously in the case of test persons voluntarily sharing their funds in the dictator game). What Fehr et al. have in mind when they call A's behavior *altruistic* seems to be something very much along these lines. As they see it, "altruism" refers to the fact that A's anticipated behavior made mutual transferal more frequent among B and C, very much to B's and C's mutual benefit (remember that the transferred sums were tripled by the experimenters, so that by both choosing to transfer, both parties ended up better off). By making unilateral non-transfer (i.e. the attempt to cash in the other's money while keeping one's own) less attractive, A's presence ultimately benefited both. While being altruistic with regard to the group consisting of B and C, however, it seems that in both a more narrow and a broader perspective, A's behavior could be called *destructive* or *aggressive* rather than *altruistic* and *pro-social.* In a more narrow perspective, it is hardly an act of altruism to inflict loss on another person; B is *the victim of A's aggression* rather than the beneficiary of A's altruism; and in a broader view, it is far from obvious why the *increase of the total cost of the experiment* that resulted from A's presence should be considered an act of altruism! After all, somebody had to cover these costs, too (presumably the taxpayers). La-

> beling *A*'s behavior as 'pro-social' and 'norm-enforcing' is as questionable as the label "second order public good", for she seems to be supporting the *appropriation of the experimental fund for the private benefit of B and C*, which, in itself, is hardly a socially desirable outcome. To this second line of argument one might object that the deliberative process of the participants is limited to the experimental situation, so that questions such as where the experimental funds came from and what the stakes of some wider public in the current situation are were not of concern to the participants. This is undoubtedly true (and it is indeed part of the whole idea of experimental games), but as an essential part of how the participants 'frame' game experimental situations, it is in itself a remarkable fact that is in need of explanation.

My claim is not that the general thrust of Fehr et al.'s interpretation of their experiment is mistaken; but whether the behavioral pattern discovered by Fehr et al. is rightly called "beneficial" or "pro-social" rather than "aggressive" or "destructive" is not self-evident from the mere *lines* of the experimental sketch, i.e. from the choices made (and the payoffs received). The problem encountered here is similar to that of the interpretation of Saint-Exupéry's 'Drawing Number One': in order to decide what sense to make of the sketch (snake or hat), we need to have the right kind of *background understanding* of the situation at hand. As to the structure of this background, I put forward three interrelated claims:

1) The structure of the behavior in question cannot be adequately described within the conceptual distinction between behavioral egoism and altruism.

2) A strong concept of group-relatedness is needed in order to make sense of the observed behavior.

3) An adequate account requires us to depart from the atomistic notion of the agents' identity as implied in much of standard economic theory.

I proceed as follows. In the next section (2.) I approach the limitations of the distinction between egoism and altruism from a historical perspective, before introducing the role of social identities in the following section (3.). In the concluding section (4.) I develop a fuller account of Fehr et al.'s experiment.

2. Beyond Egoism and Altruism

For an adequate understanding of Fehr et al.'s *third party punishment*-experiment it is crucial to understand that the interaction of *B* and *C* is a classical *prisoner's dilemma* (PD). Both are better off if both decide to transfer. However, each one is better off if she or he does *not* transfer. If both decided to transfer, each one receives three times the transferred sum. If only one transfers, the party that does not transfer cashes in three times the transferred sum in addition to his own money, which he keeps, while the transferring party ends up with nothing. If neither transfers, both keep their own money.

The PD is the best analyzed problem in all of current social theory. And yet, for all of the attention it has attracted over the last half century, a certain way of *interpreting* the PD has been so predominant that it is rarely noticed in the current debate that it is an *interpretation* of the PD rather than the PD itself. In this interpretation, the PD illustrates something like the tragedy of economic rationality, which leads to Pareto-inefficient results, or the impossibility of mutually beneficial *cooperation* among rational *egoists*. As I shall argue, however, the labels "egoist" and "cooperation", as applied to the PD, depend on a certain interpretation of the social identity of the participating agents, which has consequences for what counts as a *solution* to the problem.

The conventional interpretation is rooted deeply in the history of the PD. The original design of the game is usually credited to Merrill Flood and Melvin Dresher, whose original idea was given its name and narrative clothing (to be inspected in more detail below) by Albert W. Tucker (Tucker [1950]1980). Around the same time (early in 1950), and without any apparent connection to the former, Howard Raiffa stumbled upon the PD in his own game theoretical and game experimental research. Raiffa recounts having had no qualms whatsoever calling mutual defection the "solution" to the PD "from a descriptive and prescriptive perspective". According to Raiffa, the whole point of the PD is simply that "two stupid players do better than two smart players".[2] Raiffa's rather hard-nosed attitude to the problem of the PD seems to be typical for what one might label the 'orthodox' view. It is important to notice, however, that in this orthodox view the PD is not a *dilemma.* It is part of the definition of a practical dilemma that the agent is forced to choose between equally repellent alternatives.[3] For rational, un-sympathetic egoists such as Raiffa's 'smart players', however, there is no pondering over what to choose in the PD. There is one strictly dominant strategy and whoever is 'smart' enough to let her or his choice be determined by the expected payoff will have to choose accordingly. Within the orthodox framework the PD is no dilemma. Rather it is a practical *paradox* (as which it is indeed often referred to in the relevant literature): by each one choosing what seems to be the optimal strategy independently of the decision of the other, both participants end up worse off. In this situation, rational choice paradoxically turns out to be a rather ineffective means of maximizing one's utility.

Yet for both normative and descriptive reasons, the hard-nosed 'orthodox' attitude to the problem of the PD is met with increasing criticism. As far as the normative dimension is concerned, one might see Raiffa's alleged 'smart players'

[2] Raiffa 1992, 172; Raiffa startet having qualms only when he considered finite *repetitions* of the PD, feeling rather 'dismayed' at the prospect of constant rational non-cooperation (with correspondingly increased costs). To his relief, however, the participants in his informal experiments turned out to be more cooperative-minded (and less rational?) than he had expected.

[3] Homer provides the classical example for a practical dilemma when in the Ilias (IX/410ff.) he has Achilles pondering over his 'twofold fates', i.e. the decision whether to stay in Troy and fight or return home. Fighting, on the one hand, will earn him 'imperishable renown' which means so much in his life—but not to much avail, for his life will then be rather short. Returning home, on the other hand, will considerably prolong his lifespan, but only at the price of his renown which he values so much.

as mere 'rational fools' (Sen 1977); at least it seems overly harsh (and indeed incompatible with our pre-theoretic understanding of the term "rationality") simply to discard mutual cooperation as 'stupid' and irrational. And descriptively, experimental economists have revealed surprisingly high levels of cooperation even in one-shot PDs among anonymous participants (cf. Kagel/Roth 1995, 26ff.). Thus, for less hard-nosed economists and experimental game theorists the decisive question is: how can these cases be accounted for without dismissing the behavior as irrational?

A broader conception of the structure of human behavior offers a solution. All that is needed to see how rational subjects can find their way out of a PD is to break with the conception that human behavior be *narrowly self-regarding*. Whereas *egoism* is seen as the cause of the prisoners' problem, *altruism* is believed to be the solution. Psychologically speaking (I shall turn to the *behavioral* viewpoint shortly), people simply have to care about other people's payoffs enough so as not to be tempted to try to get the better of their partners by unilateral defection. However, there are serious doubts in the existing literature as to the range of cases in which altruism is effective in transforming PDs into games where cooperation is the rational choice. It appears that both in the case of *sympathetic* and *self-sacrificing* altruism (the first consisting in incorporating the other's utilities into one's own at a certain positive rate, the second consisting in replacing one's own utilities with the other's), cooperation will result only in some special cases. In the other cases, any of the following might happen: either altruism is simply *not strong enough* (where the rate at which the other's utilities are incorporated in one's own is too low), or altruism leads to *indifference* between the cooperative and the uncooperative strategies, or the transformation of the game even leads altruistic agents to *favor noncompliance*, which is likely to undermine cooperative stability (cf. Verbeek 2002, 86–102). In the latter case, a series of new dilemmas might follow from altruistic motivation, where this is common knowledge (cf. Tuomela 2000, ch. 10). To use Bruno Verbeek's words, altruism does not seem to be the one omnipotent 'cooperative virtue' which it is often claimed to be.

I do not want to go further into the details of this important discussion here but take another line to cast doubt on the use of a theory of *altruistic behavior* as a 'solution' to the Prisoner's Paradox. 'Altruism' might be seen as a plausible candidate for a solution of the PD only where 'egoism' is seen as the cause of the problem. This view, however, relies on a certain *interpretation* of the PD, namely that the parties affected by the outcome of PD-like situations (relative to whom the choices can be labeled as 'altruistic' or 'egoistic') are identical with those who take an active part in it. In most real-life PDs, however, this will not be the case. What is more, this condition is not even met by what might be called the *original* PD.

Remember the story Tucker invented to illustrate the problem (the original version, together with what might be the first notation of the PD beyond Dresher's blackboard, is reprinted in Tucker 1980, 101): the two players are introduced as a team of "two men, charged with a joint violation of law", and "held separately by the police". They are presented with the well-known deal.

Separately, they now face the decision whether to confess (and thereby to implicate the other). However, this is not all: interestingly, there is a *third party* present in Tucker's story, whose role is all but forgotten in later accounts of the PD. It is *the State* who—for obvious reasons—has rather high stakes in the matter. Even though the State, as Tucker observes, "exercises no choice" in the PD, it "receives payoffs" (Tucker 1980, 101). From the point of view of the State (i.e. the general public which is represented by the state), the Nash equilibrium—mutual confession—is the *optimal outcome* (the public has a vital interest in a high crime detection rate). In Tucker's payoff matrix, unilateral confession of either prisoner is the second best outcome,[4] whereas mutual non-confession is the worst outcome from the perspective of the public (the crime remains unpunished, crime detection rate is lowered).

The presence of this *third party* substantially alters the situation at hand. The prisoners' choice is not just whether to *cooperate* with each other or *defect*; it is whether to *collaborate* with each other or to *cooperate* with the third, i.e. the State, or the public.[5] In the usual game theoretical notation of the situation, the third party is simply left out. Only the two prisoners, their available strategies, and their respective payoffs matter. In this *limited view*, it might indeed appear as if "confess" was the egoistic choice, "not confess" being the altruistic alternative. In a wider perspective that includes third parties, however, the simple conceptual distinction between egoism and altruism is of little use. With just as much right as they are traditionally seen as rational egoists, the two prisoners choosing to confess and to implicate each other could be seen as rational *altruists*. The reason is obvious: while the two prisoners, by mutually confessing and implicating each other, fail to further their respective self-interests, they are quite effective in furthering the interests of the wider public by contributing to a high crime detection rate.[6]

It might be a little hard to believe, though, that their concern for the public interest is what moves the prisoners when they choose to confess (though some such cases have been reported). Yet we do not need to resort to any such motivational story if we restrict the use of the term *altruism* to its *behavioral* meaning, as it is the case in Fehr et al.'s analysis of the nature of altruism. Here, as in the entire debate on the interpretation of Fehr's experimental games, "altruism" is defined as *a costly act that confers benefits on other individuals*, regardless of the psychological background (cf., e.g., Fehr/Fischbacher 2003, 785). Even in this behavioral sense, however, mutual defection in PDs, where third parties are negatively affected by cooperative outcomes, shows all marks of *genuine altruism*: while being costly to the agents (cooperation would have left both participants better off!), it confers benefits on other individuals (e.g. the general public).

---

[4] It seems that this is dictated by Tucker's desire to have the PD transformed into a zero-sum game; in this respect, the role of the State in Tucker's PD might be seen as that of a *deus ex machina*.

[5] One might, of course, quarrel over which alternative should be labeled 'cooperation', and indeed this is precisely what this paper is about: a matter of the *determination of the social identity of the participants in question*.

[6] Apparently without knowing that this three party-setting was already part of Tucker's original conception of the PD, Elizabeth Anderson emphasized this point (Anderson 2001).

Thus the conceptual distinction between egoism and altruism is unhelpful when it comes to describing agent's choices in PDs, where third parties are affected. In particular, this concerns cases where mutually defective or cooperative outcomes result under conditions where

1) the outcome resulting from mutual individual expected utility maximization is Pareto sub-optimal with regard to group s,

2) the agents are members of s, and

3) the outcome which is Pareto-optimal for s leaves a third party t worse off than the outcome resulting from mutual individual expected utility maximization.

As far as mutual defection is concerned, the above considerations have already indicated that to some degree it might be left to one's disposal whether one likes to call the respective defective choices egoistic or altruistic: they can be *both*. Even more interesting is the case of mutual cooperation, for the according behavior seems to be *neither* of the two. Whereas in this (as in any other) case, cooperation clearly shows all the marks of altruism with regard to the other members of s—it is costly to the agent (who foregoes the benefit of unilateral defection), and it confers benefits on the other (sparing her the fate of being made the 'sucker')—it is much more difficult to account for the members of t in this story. *From their perspective* the distinction between egoism and altruism simply collapses. With regard to those unfortunate third parties, cooperation is definitely not *altruistic*, for it leaves them worse off. But it is clearly not *egoistic* either, properly speaking, for there was an alternative that would have left both the agent and the members of t better off (i.e. defection at the expense of the other members of s).

The conceptual confusion that results from a wider view that extends beyond the participating parties in a PD calls for clarification. The distinction between egoism and altruism does not work here. Apparently, it is not enough to go beyond the traditional selfish model of human behavior by allowing for altruistic behavior; in order to do justice to those pervasive cases of non-selfish behavior, where an optimal outcome for s inflicts losses on t, more structure has to be added to the behavioral picture. In the current debate, terms like "group-directedness" are introduced for such purposes (cf. Tuomela 2000). I prefer to use a neologism coined by the Spanish philosopher José Ortega y Gasset, who in his book on *Man and People* introduced the term *"nostrism"* or *"nostristic attitude"* (Ortega y Gasset 1957, 150). He coined this neologism because he understood the need to go beyond egoism and altruism in order to capture the sense in which much of our behavior is structured. 'Nostristic behavior' is neither self-directed (or egoistic) nor other-directed (or altruistic) but oriented towards our *shared goals and concerns.*

3. Nostrism: the Role of Social Identities

The claim that the analysis of group-related behavior requires more conceptual tools than the conceptual distinction between egoism and altruism is not uncontroversial. Elliot Sober's and David S. Wilson's position on the structure of unselfish behavior, for instance, does not imply any such irreducible concept. Even though in the second part of their seminal book, they propose a thoroughly 'nostristic' reading of non-selfish motivation (in accordance to which "the 'I' is defined by relating it to a 'we' "; Sober/Wilson 1998, 233), they seem to take group-directed behavior to be a *mix* of egoism and altruism rather than a third, altogether different type of attitude.

Yet the alternative between reducing group-directed attitudes to a mix of egoism and altruism on the one hand, and introducing *nostrism* as an independent third type on the other hand is not exhaustive. Perhaps nostrism is neither of the two but rather a *general structure* of which a) egoism is a *marginal case* and b) altruism is an internal feature (for all cases with the exception of a). This can be explained as follows:

a) nostrism becomes *egoism* to the degree that s is shrunk so as to contain only the agent himself.

b) nostrism implies an *altruistic attitude* towards the other members of s.

The analysis of the *mode of reasoning* underlying the *nostristic attitude* and its implications for our understanding of the structure of cooperation have become the center of an extended debate in the last two decades (cf. esp. the work of Tuomela 1995; 2000, as well as, among others, Gilbert 1989; Sugden 1993; 2000; Gold (ed.) 2005). Other lines of thought such as Amartya Sen's concern with the structure of *committed action* and the role of *identity* in his critique of rational choice theory fit seamlessly in this general venture (Schmid 2005; Peter/Schmid (eds.) 2006). Most of the work carried out in this context does not, however, *directly* pertain to the question of altruism as it arises in experimental game theory, since it is aimed at exploring the *reasons, motives* and *intentions* of the agents rather than giving functional explanations. However, no functional explanation of group-related behavior can remain completely indifferent as to the question of the possible 'proximate explanations' of such behavior. The question imposes itself as to which *motives, preferences, and modes of reasoning* can be interpreted as the most likely candidates for having evolved to sustain the respective behavior. For beings whose behavior is not exclusively prompted by instincts and immediate urges but who can think and deliberate, theories such as the one of we-mode thinking (Tuomela) team reasoning (Sugden), collective intentionality (Searle), joint commitment (Gilbert) offer the most plausible candidates.[7]

[7] Antti Saaristo argues that at least where the evolutionary basis is conceived of in group selectionist terms, collective intentionality is the most likely candidate for a proximate explanation (Saaristo 2005, ch. 2). Fehr et al. claim that group selection is not necessary for the evolution of strong reciprocity (Bowles et al. 2003).

It is an open question whether, by means of some utility transformation rule, nostristic agents' choices can be fitted into the classical game theoretical framework. I do not intend to pursue this general issue here[8] but limit myself to addressing the special question of how the presence of nostristic agents alters experimental PDs in a non-formal way. Let us define an agent's *identity* simply as her membership in (or belonging to) the group to whose optimal collective choice her individual choice contributes. Thus, an individual utility maximizer's identity consists simply in being herself, while a cooperating prisoner's identity consists in his being a part of the team of two criminals. How do these identities alter the PD? As mentioned above, the prisoner's dilemma is no dilemma for agents whose identity is limited to their individual selves but a *practical paradox* (by doing their best to maximize their individual utilities they end up being worse off). One might think this paradox does not arise for team members. As it turns out, however, this is not necessarily the case: agents whose identity extends to the group of their possible co-operators may face no less of a paradox than the individual selves. The problem is the following: as unconditional co-operators, hard-nosed team thinkers will inevitably attract *free-riders* (who can always count on unconditional cooperators). In many cases this will result in a paradoxical effect. For depending on the circumstances it might well be that unilateral cooperation resulting from the team-thinker's being abused by free-riders is *even worse an outcome from a group perspective* than mutual defection. Where there is common knowledge of this structure but not of the participant's identity, the PD becomes a paradox even for hard-nosed team-players.[9]

Thus the Prisoner's Dilemma is a *practical paradox* not only for unconditional

[8] In the respective debate there are several attempts at formalizing the group-oriented point of view within the classical game theoretical framework, e.g. by applying transformation rules (for a detailed discussion cf., Tuomela 2000, ch. 10). It seems to me, however, that Tuomela 2000, Anderson 2001, and Hurley 1989, as well as many others are right to say that there might be a systematic barrier to any such attempt. It could be this: the game theoretic framework imposes an *act consequentialist* understanding of choice (where choices are understood as the *causes* of the outcomes). It appears, however, that if people contribute to shared practices, they conceive of their individual choices not just in terms of cause and effect but in terms of *parts and wholes*, too. People who choose to cooperate in one-shot prisoner's dilemma situations choose to do *their part* in what is *better for both.* Their choices are intended as contributions constituting an optimal collective choice (which can then in turn be evaluated in act consequentialist terms). Thus it seems that a *constitutive* understanding of choice is required which defies the framework of classical decision theory.

[9] Thus, one might imagine that in some case similar to the original story of the Prisoner's Dilemma it comes as quite a shock to Row who has chosen to confess and thereby to implicate her companion when she learns that for his part Column had decided *not* to confess and not thus not to implicate Row (meaning that Row now gets off scot free while Column faces an extra long time in prison). It might well be that Row is not pleased with this result at all; she had decided to defect not because she was particularly keen on getting the better of Column. Rather, Row had reasoned as follows: while she did never suspect Column of trying to get off the situation by unilateral defection (because she knew Column to be too much of a team-player for that), Row felt certain that Column could not possibly have enough confidence in her as to choose to cooperate (which, as she is later to find out to her great distress, was a wrong belief). Because she was not particularly keen on ending up the 'sucker' herself, and expected something similar to be true of Column, she decided to defect, and expected Column to defect, too. In this situation, Column's unflinching team-mindedness paradoxically turns out to lead to an outcome that is worse *for both.*

defectors (i.e. individual utility maximizers), but *for unconditional cooperators* (i.e. hard-nosed team players), too. For neither of them, however, is there anything *dilemmatic* in this situation. Neither *homines oeconomici* nor their cooperative-minded counterparts, *homines sociologici,* will seriously be able to take both possible choices into consideration. For an individual utility maximizer, 'defect' is just as obvious a choice as 'cooperate' is for such 'over-socialized' agents as fully class-conscious workers or citizens of the ideal 'kingdom of ends'. Both images of human behavior are similarly askew. We need to go beyond either type of *a priori* fixation of human identity in order to understand what is so gripping about the Prisoner's Dilemma. It is a practical *dilemma* only for those agents whose identity is not *fixed* either to themselves or to a given group. It is a real practical *problem* only for those agents who can (and have to!) *determine* their identity making their choices.[10]

Agents whose identity is not fixed *a priori* can 'see' the situation at hand both as individual utility maximizers and as team-players, and where necessary they will make the task of having to determine their identity by making their choices a part of their reasoning. Even in a one-shot PD situation, Row (who is in a PD with Column) might start out her chain of reasoning by conceiving the situation at hand from the perspective of her individual viewpoint. Given the paradoxical effect of individual utility maximization, it soon becomes evident to her that it would be much better for both to frame the situation *as a team.* As soon as she re-conceives the situation as 'one of us', however, it immediately strikes her that given the strong pressure for unilateral defection, it is very likely that her decision to cooperate will lead to an outcome that not only leaves her with the sucker's payoff but seems even worse than mutual defection from the team perspective. Thus it becomes attractive even from a *team perspective* to adopt the 'each one for himself'-approach again and mind one's own utility, looking at the situation not as a group member but as an individual. This leads her back to the beginning of her chain of reasoning. Oscillating between her identities as an individual and as a member of the team, Row has to choose between two equally paradoxical alternatives. Having to determine their identity (i.e. having to choose the 'unit of optimization', as it were) is the prisoner's real dilemma.

4. Making Sense of Drawing Number One

What is the importance of these considerations for the interpretation of the results of Fehr et al.'s *third party punishment*-experiment? At first, the connection might not seem obvious. Compared to Row and Column in the above case, the participants in the first round of Fehr et al.'s experiment (above: B and C) found themselves in a much easier position. Apparently, they did not have to go through any such identity-shaking considerations, for they knew the Third Party (above: the test person A) to be watching them and to exert his influence

[10] It is most important not to conceive of the determination of the agent's identity in terms of a *rational choice,* for this immediately sets off an infinite regress (cf. the 'priority of identity to rational principle' in Anderson 2001).

in the experiment, and it seems that they quite correctly suspected that any unilateral defection from their part would trigger the third party's wrath (in similar experiments, surveys revealed that the participants expected the third party to 'punish' unilateral defection even more severely than he actually did). In a faint analogy to Jean-Paul Sartre's social ontology, where the third party's view 'glues together' the interacting 'I's' to a 'We', the third party here makes cooperation a *more rational choice*. This is certainly the case for those *conditional cooperators* whose identity is not fixed to themselves and who are therefore willing to do their part in what is best for both *if the other one follows suit*. More than that, the third party even seems to make any understanding of the two as *members of a team* superfluous by bringing their individual self-interest in harmony with what is best for both. Where a sufficient number of third parties can be counted on being around, even simple textbook *homines oeconomici* with their identities *a priori* limited to themselves will cooperate. Thus the result of the experiment seems to be that the *homo oeconomicus*–model of human behavior and its atomistic view of the agent's identity can be left intact, only that some 'police' has to be added to the picture. Where *third party punishers* are around, cooperative behavior is simply individual utility maximization under conditions where social norms are enforced (by sanctioning deviant behavior). No further considerations concerning the agent's identities (we-mode thinking, team reasoning, collective commitments and alike) are necessary.

I think, however, that any such reading of Fehr et al.'s results would be profoundly mistaken. A conception of social identity in terms of team membership and collective agency is needed if we are to make sense of *the third party's behavior*, and of the cooperative norms she enforces. The core idea is that in more complex settings, cooperation and social identity are mutually explicative. If one labels a certain behavior as 'cooperative' (or 'altruistic'), this is meaningful only with regard to a *certain limited set of people,* whereas one can always find some *other set of people* with regard to whom the same behavior would have to be called defective (or 'egoistic'), and *vice versa*. Thus in so far as the third party is interpreted to sanction *defection* or to enforce *cooperation*, the problem concerning the determination of the participants' identity in a PD by no means becomes an obsolete issue. Rather, it is transformed into the question of *how the identity of the participants is determined by the third*. Had she chosen the relevant social identity to be that of the *tax payers*, the third party would probably have sanctioned the appropriation of the experimental fund for private benefit which resulted from mutual transfer in the first round of the experiment.

In order to make sense of the third party's actual behavior as recorded by Fehr's experiment, one has to assume that in his perception of the situation at hand, the relevant social identity was the *team of the two participants* of the first round.[11]

It is a well-known fact that 'shared identities' are of great influence on coop-

[11] Remember that the concept of identity, as introduced above, is not a 'thick' concept. In this sense, identity does not necessarily involve such elements as a shared history and a common perception of the situation. In this sense, the total anonymity of the experimental situation does not render impossible the emergence of shared identities.

eration in social dilemma situations. In *social identity theory* as well as in other research programs, experimental studies have repeatedly shown that cooperation rates between members of the same group are much higher than between members of different (or even competing) groups, where the participants know about their partner's social identity (c.f., e.g., Kollock 1998). But how do social identities arise? How are social identities determined? How does the perception of the situation as a member of one or another group become *salient* to the participants in the experiment, especially where anonymity is part of the experimental setting? In the case of the experimental game at hand it seems plausible that the *instruction given to the participants by the experimenters* might have played an important role.

That the 'principle of description invariance' (according to which it should not matter to the outcome how an experiment is described) does not hold is nothing new (Tversky/Kahnemann 1986). Camerer and Fehr, however, suggest that in their experimental games such effects are minimized by avoiding 'concrete' descriptions.

> "The games are usually described in plain, abstract language, using letters or numbers to represent strategies rather than concrete descriptions like 'helping to clean up the park' or 'trusting somebody in a faraway place'. As with other design features, abstract language is used not because it is lifelike, but as a benchmark against which the effects of more concrete descriptions can be measured." (Camerer/Fehr 2004, 58)

A closer look at the instructions given to the participants of the third party punishment-experiment,[12] however, reveals that at least as far as the question of *social identity* is concerned, the description of the experiment was *much more concrete* than these standards seem to suggest. Firstly, the two parties in round one are explicitly introduced as a *'group'* and are repeatedly referred to as 'team members' throughout the instructions. In line with this labeling, the possible results of the first round of the game are explained in a list with the respective payoffs labeled 'your income' and 'income of the other *member of your group*', respectively. These instructions were known to the third party (who had been actively involved in the first round of an experiment as a member of another group). Thus it seems quite understandable that she came to determine the *team of the two* (rather than, say, the taxpayers, or a team consisting of herself and the experimenters) as the relevant social identity underlying the 'cooperative norm' which she decided to enforce in the second round.

How, then, should the third party's behavior be labeled? In behavioral terms it is not *egoistic* or *self-directed.* Is it therefore *other-directed,* as Fehr et al.'s label 'altruistic' suggests? Should the result of the experiment be taken to show that human beings do not like having defectors around (or rather: like to inflict losses on them wherever they find them), regardless of how the social identity relative to which the behavior in question appears as 'defection' relates to their

[12] I wish to thank Ernst Fehr and Urs Fischbacher for giving me access to this material.

own social identity? I think that this is rather implausible, and in the experiment at hand it is clearly not the case. The set of people whose total payoff the third party's behavior optimized was the *group of all participants of the experiment*, including herself. Again, this particular social identity is heavily supported by the instructions given by the experimenters; the label "participant" (which is repeatedly used in the instructions) alone makes the total set of interacting individuals salient in terms of social identity. In addition to that the participants are told that they are taking part in an *experiment.* The term *experiment*—and even more clearly the term *experimental game*—again suggests and supports an understanding of the situation at hand according to which no 'outside relations' matter. Just like a game, an experiment presupposes some sort of *isolation* from outside influence. Thus individuals who are told that they are *participants in the experiment* will almost inevitably see themselves as 'one of the participants' rather than, say, family members, or taxpayers. Thus, the third party's behavior should be seen as a sign of a 'nostristic' attitude as one of the participants of the experiment, i.e. a personal investment in the maintenance of the 'normative infrastructure' that is 'best for us' in terms of the total group of the participants in the experiment.

This nostristic reading of Fehr et al.'s experiment does not conflict with Fehr et al.'s own proximate explanation, according to which A's behavior is driven by his *emotional response* to the observed behavior. It is a well-known fact that our emotional responses are heavily influenced by our perception of the social identities of the participating parties. We tend to respond in different ways to cases of unilateral defection depending on whether or not the defecting party or his 'victim' is 'one of us'. Nostrism is not just a matter of cognition; it is a matter of affection, too. More than that, the nostristic reading of Fehr et al.'s results seem to be in tune with most of the labels which Fehr et al. attach to their results. If we conceive of the social identity of the group of participants as relevant in the situation, most of Fehr et al.'s abovementioned labels make sense. The third party's behavior does indeed appear to be 'pro-social' and 'norm enforcing', as Fehr et al. claim—even though these terms should be handled with care because they make only sense on the basis of a prior correct identification of the relevant social identity. Something similar is true for the label "reciprocity". *In terms of personal interaction,* the view put forward above in section i. might be correct: because of the structure of the experiment there was no reciprocity whatsoever involved. However, it is not absurd to call the third party's behavior reciprocal *in the sense that she is doing her part in a communal cooperative practice.* The reciprocal element in her behavior consists in the fact that she sanctions defective behavior in the team over which she watches *just as she can count on being watched by another 'third party'* in her own first round of the experiment. The fact that the person she can punish is not identical with the person by whom she can be punished herself does not mean that the attitude is not one of 'reciprocation' in terms of *doing her part within the grid of interchangeable roles.* One might still quarrel over whether or not 'punishment' is a lucky terminological choice for the behavior at

hand. In any case: in order to make good sense of the experiment, it is essential to understand the *social identities of the participating agents.*

The experimental games show *how little* it takes for people to come to conceive of themselves (and act) as members of a team. In view of the result of experimental economics, the Aristotelian *dictum* of the *zoon politikon* takes on a new meaning: even with complete strangers whose names and personal identities they do not know, and even in pure one-shot interactions, people tend to see themselves (and conceive of others) *as members of teams.* Contrary to what one of the fathers of the PD (Flood 1958, 12) has conjectured, cooperation seems to be much less a matter of some *social relationship* between the players (there is none), than a matter of their *social identity.*

Thus, it appears that the lesson to be learned from game experiments is not just how far the orthodox economic assumption of narrow self-interest is from reality. Such experiments shake yet another pillar of economic theory, the last pillar that has largely survived the recent boom in reconsideration and reconceptualization of the economic model of human behavior. What is at stake here is the *methodologically individualist* view that social phenomena should ultimately be explained exclusively in terms of *individual action*, without there being any *collectivity concepts* involved at the basic level of explanation. In cases such as the one at hand, however, the agent's identity has to be determined before it is even possible to make sense of the observed behavior. Thus there is no 'understanding' of the behavior of individuals without *first determining the relevant groups* of which these individuals are members. Thus groups are not 'secondary' to (or 'supervenient' on) individual action but an essential feature of the most basic level of social reality.

Bibliography

Anderson, E. (2001), Unstrapping the Straitjacket of 'Preference': a Comment on Amartya Sen's Contributions to Philosophy and Economics, in: *Economics and Philosophy 17*, 21–38

Bowles, S./E. Fehr/H. Gintis (2003), Strong Reciprocity May Evolve With or Without Group Selection, in: *Theoretical Primatology Project Newsletter 1/12*

Camerer, C. F./E. Fehr (2004), Measuring Social Norms and Preferences. Using Experimental Games: A Guide for Social Scientists, in: J. Henrich et al. (eds.), *Foundation of Human Sociality. Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford

Fehr, E./S. Gächter (2002), Altruistic Punishment in Humans, in: *Nature 415*, 137–140

— /U. Fischbacher (2003), The Nature of Human Altruism, in: *Nature 425*, 785–791

— / — (2004), Third-Party Punishment and Social Norms, in: *Evolution of Human Behavior 25*, 63–87

Flood, M. M. (1958), Some Experimental Games, in: *Management Science 5*, 5–26

Gilbert, M. (1989), *On Social Facts*, Princeton

Gold, N. (ed.) (2004), *Teamwork. Multi-Disciplinary Perspectives*, New York

Henrich, J. et al. (eds.) (2004), *Foundation of Human Sociality. Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford

Hurley, S. L. (1989), *Natural Reasons*, Oxford

Kagel, J. K./A. E. Roth (eds.) (1995), *The Handbook of Experimental Economics*, Princeton

Kollock, P. (1998), Transforming Social Dilemmas: Group Identity and Co-operation. In Peter A. Danielson (ed.), *Modeling Rationality, Morality, and Evolution*, New York, 185–209

Ortega y Gasset, J. (1957), *Man and People*, New York

Peter, F./H. B. Schmid (eds.) (2006), *Rationality and Commitment*, Oxford

Raiffa, H. (1992), Game Theory at the University of Michigan 1948-1952, in: E. Roy Weintraub (ed.), *Toward a History of Game Theory*, Durham, 165–175

Saaristo, A. J. (2005), *Social Ontology and Agency. Methodological Holism Naturalised*, unpublished manuscript (doctoral dissertation to be presented to the University of London)

Schmid, H. B. (2005), Beyond Self Goal Choice: Amartya Sen's Analysis of the Structure of Commitment and the Role of Shared Desires, in: *Economics and Philosophy 21*, 51–63

Sen, A. K. (1977), Rational Fools. A Critique of the Behavioral Foundations of Economic Theory, in: *Philosophy & Public Affairs 6*, 317–344

Sober, E./D. S. Wilson (1998), *Unto Others. The Evolution and Psychology of Unselfish Behavior*, Cambridge/MA

Sugden, R. (1993), Thinking as a Team. Towards an Explanation of Nonselfish Behaviour, in: *Social Philosophy and Policy 10*, 69–89

— (2000), Team Preferences, in: *Economics and Philosophy 16*, 175–204

Tucker, A. W. (1980), On Jargon: The Prisoner's Dilemma. A Two Person Dilemma, in: *UAMP Journal 1*, 101

Tuomela, R. (1995), *The Importance of Us. A Philosophical Study of Basic Social Notions*, Stanford

— (2000), *Cooperation. A Philosophical Study*, Dordrecht

Tversky, A./D. Kahnemann (1986), Rational Choice and the Framing of Decisions, in: *Journal of Business 59*, 251–278

Verbeek, B. (2002), *Instrumental Rationality and Moral Philosophy. An Essay on the Virtues of Cooperation*, Dordrecht