Analyse & Kritik 27/2005 (© Lucius & Lucius, Stuttgart) 197–211

Jon Elster

Fehr on Altruism, Emotion, and Norms

Abstract: I discuss recent work by Ernst Fehr and his collaborators on cooperation and reciprocity. (i) Their work demonstrates conclusively the reality and importance of non-self-interested motivations. (ii) It allows for a useful distinction between trust and blind trust. (iii) It points to a category of quasi-moral norms, distinct both from social norms and moral norms. (iv) It demonstrates how social interactions can generate irrational belief formation. (v) It shows the potential of punishment for sustaining social norms and for overcoming the second-order free rider problem as well as obstacles to group selection. (vi) It offers a provocative experimental basis for the 'warm-glow' explanation of altruistic behavior. I conclude by suggesting some experiments that might allow for further developments of the theory.

1. Introduction

The research program of Ernst Fehr and his several collaborators (hereafter referred to collectively as "Fehr") is having a considerable impact on economics, psychology, evolutionary biology and neurophysiology. (I limit myself to the eight articles listed in the references). It also has the potential for clarifying important issues in sociology and moral philosophy. In elegant and tightly controlled experiments Fehr has not only demonstrated the reality and importance of altruism, trust, norms and reciprocity, but clarified the nature of the concepts themselves. Much work remains to be done. The extension of the paradigm to non-experimental settings will require considerable ingenuity, as some of the crucial features of the experiments, notably the anonymity of the agents to one another, will typically be lacking in real-life contexts. The implications for evolutionary biology are, for the moment, mainly speculative. This being said, the steady stream of publications issuing from his team are already changing the way we look at the social and natural world.

Most of the experiments are organized around *four games* in which subjects can choose between cooperative and non-cooperative, or generous and ungenerous, behavior. (For the sake of brevity, I shall refer to both cooperative and generous choices as 'cooperating' and to both non-cooperative and ungenerous choices as 'defecting'.) First, there is the simultaneous-decision Prisoners' Dilemma (PD), in which subjects choose whether to make a costly contribution to a public good from which everyone, contributors and non-contributors, will benefit. Second, there is a sequential PD, or 'trust game' (TG), in which one player has the option of transferring a sum of money, which is then multiplied by the experimenter, to another who then has the option of transferring some

of the gains back to the first player. Third, there is the Ultimatum Game (UG), in which one player proposes a division of a sum of money between himself and another player, and the second can either accept the proposal and receive his proposed share or ensure that neither gets anything by rejecting it. Fourth, there is the Dictator Game (DG), in which the Proposer unilaterally imposes a division of the sum. To these standard games Fehr adds an important twist, by sometimes allowing subjects to punish others, usually at some cost to themselves, if others behave ungenerously or uncooperatively.

2. Beyond Rational Self-Interest

Fehr's research program is acquiring a momentum that is comparable to that of the work of Daniel Kahneman and Amos Tversky in the 1970s and 1980s. Substantively, though, the two research agendas are very different. To simplify, Kahneman and Tversky brought about a revolution in the way we think about cognition, whereas Fehr is changing how we think about motivation. To be sure, Kahneman and Tversky also engaged with issues of motivation (loss aversion), but their work in this area retains a distinct cognitive or 'cold' flavor. Fehr, by contrast, addresses core 'hot' motivational issues, notably emotions and social norms. The joint effect of the two efforts is to *undermine the paradigm of rational self-interest*, Kahneman and Tversky by demonstrating pervasive violations of rationality and Fehr by showing extensive deviations from self-interest. In Section 8 I shall briefly ask whether the two agendas might be joined.

The problem of identifying non-self-interested motivations in behavior has been complicated by the fact that self-interested agents might be induced to *mimic* altruistic behavior (Elster 2004a). If sufficiently farsighted, they may find cooperation in an iterated PD to be in their self-interest. They may behave cooperatively or generously in order to build up a useful reputation for altruism, or seek to achieve the same aim by engaging in apparently pointless punishment of non-cooperators. When the rich give generously, it may be from fear of being ostracized if they didn't, or from a desire to humiliate the less generous among their peers. Economists have shown endless ingenuity in coming up with reductionist arguments of this kind.

Fehr's strategy for rebutting reductionism is to demonstrate altruistic behavior under conditions that eliminate interactive mechanisms of the sort described in the previous paragraph. (As we shall see in Section 7, the strategy does not eliminate all forms of reductionism.) The two key features of the experiments are *anonymity* and *one-shot interaction*. Since subjects communicate only through computer terminals, there is no occasion for face-to-face interaction that might generate conformist pressures. In several experiments, there is even experimenter anonymity, in the sense that subjects are told, credibly, that the experimenter will not be able to identify choices at the individual level. Moreover, in many of the experiments subjects who interact in one game know that they are not going to play against each other in the future. Although the experiment may require subjects to play the same game many times in succes-

sion, they do not play it against the same partners. They do not, therefore, have an opportunity to punish defectors in order to benefit from the cooperation the punishment might induce in subsequent games. Third parties might, however, benefit from the punishment. This is what Fehr refers to as "altruistic punishment" (Fehr/Gächter 2002). A punishes B for defecting in a game with A with the consequence that in a later game B behaves more cooperatively towards C than he would otherwise have done. The punishment is altruistic, at least in its consequences, since A's infliction of punishment on B is both costly to A and beneficial to C. Whether it is also altruistic in a psychological sense is another matter, which will concern us in Section 7.

The experiments show consistent and high levels of altruism even under these stringent conditions. Let me just cite one example among very many. In a PD without punishment opportunities, where rational self-interested individuals would make zero contributions, moderate contributions are nevertheless observed. With iteration, contributions fall but remain positive. In a PD with punishment opportunities, where rational self-interested individuals would make no contributions and impose no punishment, large contributions are observed. With iteration, contributions increase. Since no subject ever met any other subject more than once, there was no incentive to use punishment as a reputation-building device.

As Fehr repeatedly argues, altruism is not a universal feature of human nature, since many subjects show themselves to be consistently selfish. In the interaction between selfish and altruistic subjects, the structure of the game may either induce the former to mimic the latter (Fehr/Gächter 2002, 137; Fehr/Fischbacher 2002, C4) or the latter to mimic the former (Fehr/Fischbacher 2003, 787). From the point of view of institution-building, it is of course highly desirable to structure interactions so that the selfish have an incentive to mimic the altruists.

3. Trust

Trust, one of the most elusive concepts in the social sciences, can be elucidated by Fehr/Rockenbach's study (2003) of the negative impact of punishment on cooperation. The study is also important in that it demonstrates how instrumental punishment can have very different consequences from the spontaneous punishment that I consider in 6 below.

The experiment in question is a TG that is played in two conditions. In both, the 'investor' has the option of transferring anywhere between 0 and ten of their endowment of ten monetary units (MU) to a 'trustee'. The experimenter then triples any amount sent, so that if the investor sends 10 the trustee receives 30. The trustee can decide to transfer any amount from 0 to the whole augmented sum (30 in the same example) back to the investor. Finally, if the investor decides to make a transfer he also has to state the amount he wishes the trustee to transfer back to him.

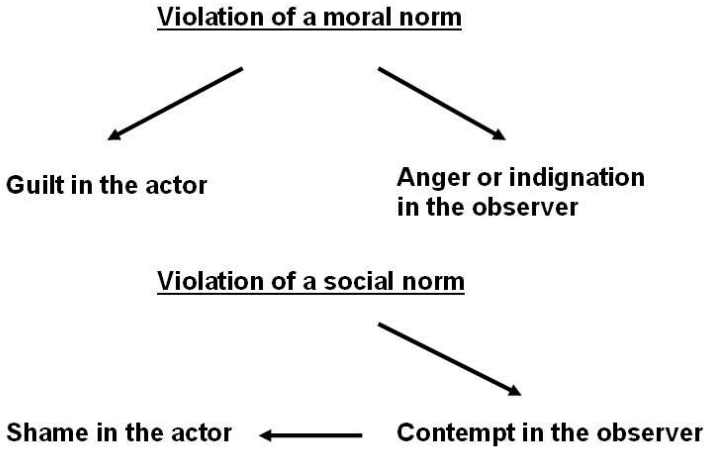
These features define the inappropriately labeled (as I shall argue) 'trust

condition'. In the 'incentive condition', the investor is also given the option of stating, at the time he makes the transfer and announces the desired back transfer, that he will impose a fine of 4 MU on the trustee if he transfers back less than the desired amount. Some investors use this option, others do not. In the latter case, trustees know that the investor had the option but refrained from using it. The results are that the largest transfers are made in the incentive condition when no fine is imposed and the smallest in the same condition when a fine is imposed, transfers in the trust condition being at an intermediate level.

One can of course define trust any way one wants. I believe, however, that the definition that best captures everyday usage may be that of *refraining from taking precautions* even when the other person might act in a way that would justify precaution. This idea of trust as lowering one's guards corresponds exactly to the incentive condition when no fine is imposed. What Fehr calls the 'trust condition' corresponds to *blind trust*, shown in situations when monitoring or punishment is impossible. The finding that (non-blind) trust induces more cooperation than blind trust combines ex-ante unexpectedness with ex-post obviousness in the way characteristic of much good research. The finding that threats of punishment can undermine cooperation is perhaps less surprising. Like trust, distrust can have a certain self-fulfilling quality. As Proust (1954, 150) noted, "once jealousy has been discovered, the person who is its object views it as a challenge that authorizes the infidelity".

4. Social and Moral Norms

Let me begin with sketching my own view of norms. They are of two kinds: social norms and moral norms. To distinguish between them, consider first the causal structure of responses to norm-violations:



The two types of norm also have different substantive content. Moral norms include the norm to keep promises, the norm to tell the truth, the norm to help others in distress; and so on. Social norms include norms of etiquette, norms of revenge, norms of reciprocity, norms of fairness, norms of equality, and so on. Some norms, about which more later, have features in common with both moral and social norms.

Violations of a moral norm can trigger either anger or indignation (Descartes 1985, Art. 195), depending on whether the observer is also a victim of the violation or a neutral third party. Through his pioneering work on second-party and third-party punishment, Fehr has put this distinction on a firm empirical basis. He does not focus directly on the emotions themselves, however, but on the *action tendencies* they trigger. All the emotions cited in Fig.1 do in fact induce a tendency to act that, if unchecked, will lead to action. Anger and indignation both induce a tendency to punish the norm-violator, but the tendency is stronger for anger. (For evidence from transitional justice supporting this idea, see Bass 2000, 8, and Elster 2004b, 232.)

The action tendency of contempt is avoidance or ostracism, i.e. the refusal to have dealings with the norm-violator. Fehr (Fehr/Fischbacher 2002, 17) cites this, too, as a form of punishment. It is true that for the victim of contempt, the avoidance can impose horrible suffering. Thus A. O. Lovejoy (1961, 181, 191, 199) quotes Voltaire as saying that "To be an object of contempt to those with whom one lives is a thing that none has ever been, or ever will be, able to endure. It is perhaps the greatest check which nature has placed upon men's injustice", Adam Smith that "Compared with the contempt of mankind, all other evils are easily supported", and John Adams that "The desire of esteem is as real a want of nature as hunger; and the neglect and contempt of the world as severe a pain as gout and stone". In material terms, too, being ostracized can be painful if others refuse to deal with you.

Such avoidance behavior may or may not be costly for the avoider. On the one hand, refusing to interact with a person because he violates a social norm might block the opportunity for a mutually profitable transaction. Novels often refer to 'gentlemen' who lose out because they refuse to deal with 'upstarts'. As Becker (1957) argued, a 'taste for discrimination' can be quite costly. School children may reject 'nerds' even when the latter could help them with their homework. On the other hand, under conditions of uncertainty, avoiding to deal with someone who doesn't wear proper attire may actually be beneficial if such behavior signals that the person is a 'bad type' who is liable to cheat, break promises, sell products of substandard quality etc. For that very reason, however, the avoidance cannot (*pace* Posner 2000, 25–26) *also* serve as a costly signal to indicate that one belongs to a 'good' type.

Following Fehr, let us assume that avoidance behaviors are in fact costly for both agent and target. In many cases this assumption is quite plausible. These situations involve three distinct costs: the material cost incurred by the ostracizer, the material cost incurred by the person who is ostracized, and the emotional suffering of the latter. Almost by definition, the intensity of the suffering caused by shame increases with the intensity of contempt. Plausibly,

the best measure of the intensity of contempt is how much the ostracizer is willing to give up of material benefits. (This proposition parallels the one that Fehr makes with regard to intensity of anger or indignation.) It follows that *the ostracized is worse off emotionally the more the ostracizer makes himself worse off materially.* If we accept the views of the authorities cited above, it also follows that from the point of view of the ostracized, the material loss he might suffer from the avoidance behavior matters less than the loss incurred by the ostracizer. What can be more humiliating than seeing that another person is willing to incur considerable expenses in order not to have to deal with me?

Contempt and anger (or indignation), I have argued, are triggered by the violation of respectively social and moral norms. By contrast, Fehr connects anger with the violation of (what he calls) social norms. This may be a substantive disagreement or a terminological one. I believe it is mainly the latter, but there are some complications. I cited as an example of a moral norm: Help others in distress. One of Fehr's paradigmatic social norms is that of helping those who have helped you or, more generally, those who have helped you *or others.* (Although to my knowledge Fehr does not refer to third-party rewards, they are implicit in his theoretical scheme.) This *norm of strong reciprocity* (Fehr/Fischbacher 2003) also, in fact especially, applies to punishments. Another paradigmatic norm is the *norm of conditional cooperation* (Fehr/Fischbacher 2004b, 186), which obliges agents to contribute to the production of public goods if and only if others or most others do so. The "only if" might be taken weakly, as asserting that non-contribution is acceptable if others do not contribute, or strongly, as asserting that it is mandatory.

It is not clear to me that these are moral norms in the common-sense meaning of that phrase. The norm of reciprocity allows you not to help others in distress unless they have helped you previously. In Fehr's scheme, C's indignation is appropriate if B does not help A who has helped B previously. In my book, C's indignation would be more appropriate if B fails to help A in distress simply because there is no prior history of assistance. The norm of conditional cooperation obliges you to donate to charity when it does little good (because others are already doing much), but allows you to abstain when it would do a great deal of good (because others are doing little). Morally, a utilitarian norm, creating a stronger obligation to give if others give little, seems more appropriate. In both cases, social norms make the appropriate action depend on what others have done, whereas moral norms make it depend on what I can achieve by acting. I am not arguing that common-sense morality always is consequentialist, only that it must be capable of identifying a *concept* of right action that is independent of whether others are acting rightly.

Social norms occupy, as it were, the space between raw, pre-social emotions and moral norms proper. On the one hand, social norms represent 'behavioral standards' that differ from spontaneous emotions of gratitude or vindictiveness (Fehr/Fischbacher 2004b, 65). On the other hand, as I have argued, social norms are too weak to generate many behaviors that we intuitively view as morally required. This being said, the norm of strong reciprocity and the norm of conditional cooperation do not have the arbitrary and sometimes dysfunctional

character of norms of etiquette or egalitarian norms that forbid anyone to have what not everyone can have. To distinguish these two norms from social norms, we may call them 'quasi-moral' norms. Because of their ambiguous status, they can be enforced by anger (or indignation), triggering punishment, as well as by contempt, triggering ostracism.

An important respect in which quasi-moral norms are close to moral norms is that people abide by them even when they are not observed by others. Moral norms are often unconditional, in two distinct senses. First, they are not conditional on others observing what the agent is doing, and second, they are not conditional on the agent observing what others are doing. Social norms are conditional in the former sense; quasi-moral norms in the latter sense. Consider the following cases:

> (i) The city authorities try to reduce water consumption by getting households to use less water. Since others cannot observe how much water I consume, I cannot be shamed into using less. Since I cannot observe how much others use, there is no scope for conditional cooperation.
>
> (ii) A person walking in the park may abstain from littering either because he sees that others don't litter or because he knows others could see him if he did. The first observation might trigger conditional cooperation, the second the operation of social norms.
>
> (iii) Imagine case (i), with the difference that the authorities regularly announce aggregate water consumption on TV. (This has actually been done in Bogotá, under the imaginative mayorship of Antanas Mockus.) Although there is still no room for social norms, conditional cooperation may be maintained since each agent can observe whether most others are doing their share.
>
> (iv) Imagine case (ii), with the difference that the person is too shortsighted to see whether others are littering although he can see well enough to notice their presence. He might be influenced by social norms, but not by conditional cooperation.

Case (iii) is similar to one of Fehr's experiments (Fehr/Fischbacher 2003, 787), in which the members of a group decide whether to cooperate or defect in one round of an iterated game on the basis of average group behavior in the previous round. The finding is that cooperation tends to unravel over time, since any level of cooperation short of 100% induces some cooperators to defect, thus reducing the level even more and inducing even more defections until some stable, low but non-zero level is reached. In my own work (Elster 1989, 204) I have argued for the opposite effect—*snowballing* rather than *unraveling*. In a heterogeneous population, the norm of conditional cooperation may be triggered by different proportions of cooperation in the rest of the population. Assume a hard core of unconditional cooperators, making up (say) 10% of the population. In the first round, they are the only ones who cooperate. In the next round, they are joined by the (say) 5% of the population who need at least 10% of the

population to cooperate before they are willing to join. In the third step, they are joined by those who are willing to cooperate at a threshold of 15%, and so on. Depending on the distribution of thresholds in the population, the process could move all the way to universal cooperation or stop short of it. A real-life example of snowballing was observed in the build-up of crowds on successive Sundays in Leipzig prior to the demolition of the Berlin Wall in 1989 (Petersen 2001, 262–69). Whether any given case will result in unraveling or snowballing depends on initial beliefs and on the distribution of thresholds.

5. Irrational Belief Formation

Although Fehr's experiments mainly aim at explaining behavior, they also throw interesting light on the *beliefs* of the subjects. Three findings stand out. First, "third-party punishment reduced the income of a defector paired with a cooperator by ... 10.05 points" but if both actors in the PD defected "each defector received on average only 0.583 deduction points ... indicating that third parties perceive the same action—in our case defection—very differently depending on what the other player in he PD did" (Fehr/Fischbacher 2004b, 73; see also Fehr/Fischbacher 2004a, 186). This implies that *third parties punish defectors on the basis of events outside their control*, since in a PD with simultaneous choices the players have no way of knowing whether their partners cooperated or not. In the real world, many PDs are sequential games with full information (as in the TG): one partner first decides whether to cooperate or not, and if he chooses cooperation the other has to decide whether to cooperate or to exploit him by defecting. In such cases, it makes sense for third parties to blame a defector who exploits a cooperating partner. It makes no moral sense, however, to blame a defector simply because unbeknownst to himself he happens to be matched with a cooperating partner. When people do so, their response might be an instance of a phenomenon that according to Fehr (Fehr/Henrich 2003) has not been shown to exist, namely that people in the laboratory react in ways that make sense only outside of it. We should note, however, that in another experiment (de Quervain et al. 2004, 1256) subjects did *not* punish those who failed to reciprocate when the failure was due to factors outside their control.

The second and third findings suggest that in some of these experiments subjects are prone to irrational belief formation. I take rational belief formation to imply the absence of systematic bias. For instance, free-riders in the PD should on average have correct expectations about the anger that cooperators feel towards them. It turns out, however, that the "anger that was expected by the free riders ... was even greater than the actually expressed anger" (Fehr/Gächter 2002, 139). Also, those who are on the receiving end of an ungenerous Dictator should on average have correct expectations about the tendency of third parties to punish the Dictator. It turns out, however, that "the proportion of recipients ... who expected [the third party] to punish was higher than the proportion who actually did so" (Fehr/Fischbacher 2004b, 68). Moreover, the tendency to overestimate the tendency to punish is a decreasing function of the dictator's

transfer to the recipient (Fehr/Fischbacher 2004b, 69). In other words, both wrongdoers and victims tend to overestimate the likelihood that wrongdoers will be punished. Conjecturally, this cognitive bias might be due to agent-specific emotions—guilt and fear in wrongdoers, anger in victims.

6. Punishment

In Fehr's recent work, punishing behavior has a central role in generating and sustaining cooperation. Let me first comment on appropriate versus inappropriate punishment, and then move on to the mechanisms by which (appropriate) punishment is capable of shaping behavior.

Punishments are subject to two frequently violated moral constraints: moral standing and relevance. First, do not punish someone for doing what you have done yourself (don't throw stones in a glasshouse). Defectors in a PD should not punish other defectors. Second, do not punish another person because of harm done to you by a third party (you should not get mad at your spouse because you got a dressing-down from the boss). In Fehr's experiments, subjects violated the first constraint but, perhaps surprisingly, not the second. He found that even defectors tend to punish defectors (Fehr/Fischbacher 2004b, 83), matching the real-life finding that those who do little to oppose a harsh political regime are often among the most eager to punish its agents when it falls (Elster 2004b, 241-43). In that case, the underlying mechanism is probably some kind of desire for redemption, as if post-transition aggression towards wrongdoers could make up for pre-transition passivity. Not inconsistently with that idea, Fehr (Fehr/Fischbacher 2004b, 83, note 6) suggests that in his experiments self-serving bias might be at work. By contrast, he found that "Third-party punishment was not significantly affected by the sum one received from one's own dictator" (Fehr/Fischbacher 2004b, 81). This is surprising in the light not only of preanalytical intuitions, but also of Fehr's own argument (Fehr/Fischbacher 2004b, 80) that anger is a function of (among other things) the level of arousal. Presumably, being the target of bad treatment by one's own dictator generates arousal that could trigger harsher punishment of the "outside dictator". The classical (although poorly replicated) findings by Schachter and Singer (1962) are relevant here.

There are two reasons why reliance on punishment might seem insufficient to sustain much cooperation. On the one hand, to the extent that punishment takes the form of ostracism and avoidance, it may not matter much to the defector. He may not be able to take advantage of a particular victim again, but there are plenty of other victims he can exploit. As the saying goes, there's a sucker born every minute. On the other hand, to the extent that punishment is costly, it is not clear why people would be willing to engage in it. What's in it for them? In a public goods situation, people are tempted to free ride by imposing negative externalities on each other. If the solution requires their willingness to impose positive externalities on each other (by punishing free riders), isn't the original

problem just recreated at a higher level? This conundrum is usually referred to as 'the second-order free-rider problem'.

Fehr's work suggests answers to both questions. His answer to the second is somewhat ambiguous, although I believe the main thrust of his argument is clear enough. Let me begin, however, with the first question. Fehr's answer (briefly stated in Fehr/Fischbacher 2004b, 85), is that third-party punishment can provide a powerful supplement to second-party punishment. Suppose that A defects in an interaction with B, in the presence of C, D ... etc. Although A's loss from B's ostracism of him may be small compared to the gains from the defection, the sum-total of the costs that follow from being ostracized by C, D ... etc. may exceed those gains.

Obviously, the physical presence of C, D ... etc. on the scene is not necessary; what matters is that they somehow gain knowledge about B's defection. I believe this is the proper perspective in which to see the relation between *gossip* and social norms: gossip acts as a *multiplier* on punishment. Suppose A engages in norm-violating behavior towards B, e.g. by letting his cattle trespass on B's land (Ellickson 1991), and that B tells C, D ... etc. about the trespass. The diffusion of information about the trespass adds a group of potential third-party punishers (C, D ...) to the original second-party punisher (B). In fact, B may deliberately pass on information about A's trespass in order to increase the social pressure on A to reform his behavior. Knowing this, A may be deterred from defecting in the first place. Following other writers on norms (Coleman 1990, 285; Ellickson 1991, 173), Fehr (Fehr/Fischbacher 2002, C18) asserts that gossiping is *costly* for the gossiper. This seems wrong. Casual observation suggests that people gossip because of the direct benefits it provides—it's *fun*. Yet even if gossip were in fact costly, Fehr has the conceptual wherewithal to explain the willingness to incur sanctioning costs, as I now proceed to show. Turning now to the second of the two questions I identified three paragraphs above, Fehr offers two responses. One is that cooperation is more likely to emerge and persist "when punishment of non-cooperators *and non-punishers* is possible" (Fehr/Fischbacher 2003, 790; my italics). This appeal to the punishment of non-punishers seems, however, empirically unfounded and theoretically weak. Empirically, the phenomenon seems to be rare. In everyday social interactions people rarely shun those who do not shun defectors. If we look at the violation of social norms rather than of quasi-moral norms of cooperation, things are different. Horrible phrases such as "Jew-lover" and "nigger-lover" remind us that people may indeed be deemed guilty by association. In societies with strong norms of revenge a person who fails to ostracize someone who fails to take revenge may himself be ostracized. Yet note that the occasion for revenge in these societies is the violation of personal honor rather than free-riding. Theoretically, the appeal to the punishment of non-punishers seems arbitrary, for who would punish the non-punishers of non-punishers? It is not only more parsimonious but empirically more plausible to argue that people *spontaneously* punish defectors. This is, I believe, Fehr's main line of argument. The bulk of his work is indeed oriented to showing that even when completely unobserved and hence invulnerable to sanctions, people are willing to spend resources on punishing defectors. This spontaneous

willingness to punish offers, as noted a simple solution to the second-order free-rider problem. It may even offer a plausible mechanism for the operation of *group selection* (Fehr/Fischbacher 2004a, 189). The main objection to group selection, namely the vulnerability of a population of cooperators to invasion by free riders, would cease to hold if the latter were liable to be punished for attempts to exploit the cooperators.

7. The Warm Glow

A common objection to the claim that people sometimes act to promote the welfare of others at expense of their own is that when they appear to do so they are in reality acting to produce the warm glow they expect to derive from benefiting others. The idea is not merely that people feel a warm glow when they benefit others, which might arise even when they act from duty, but that production of the glow is the motivating *aim* of the action. To have any content, the objection would presuppose that one can (i) measure the warm glow and (ii) show that production of the glow is the aim and not merely a side effect of the action. Until recently, these presuppositions were so obviously unrealizable that the objection could hardly be seen as anything but a piece of dogmatic cynicism. With the advent of brain scans that allow one to measure the flux of psychic reward in real time, the situation has changed. In a recent experiment (de Quervain et al. 2004), Fehr uses brain scanning to argue for a distinction between biological and psychological altruism. Even when we observe altruistic behavior in a one-shot experiment with full anonymity, the underlying motivation can be the expectation of a measurable warm glow.

In the experiment, subjects who were treated ungenerously in a TG had the option of imposing a punishment on their partner. In one condition, the punishment was costless for the punisher; in another, it was costly. In both conditions, subjects were asked to think intensely for one minute about the punishment (if any) to be imposed, during which period their brain was scanned to detect activation of brain circuits. In both conditions there was a correlation between activation of reward-related circuits and the actual monetary punishment. To distinguish between two interpretations of these findings, (i) that the decision to punish induces satisfaction and (ii) that the expected satisfaction from punishment induces the decision to punish, the authors considered 11 subjects who imposed the maximal feasible punishment in the 'costless' condition. Among these subjects, those whose reward circuits were more highly activated also imposed more severe punishments in the 'costly' condition. This finding support (ii). 'Biological altruists' punish defectors because they expect it will make them feel good, not because they want to benefit others.

The finding is striking, but how well does it generalize? Second-party altruistic punishment is altruistic in its consequences, but motivationally it is closer to revenge. It remains to be seen whether third-party altruistic punishment, second-party reward and third-party reward exhibit the same pattern. Presumably, behavior in these situations could be studied by the same brain

scan techniques, to see whether benevolence (reward behavior) and impartiality (third-party behavior) also lend themselves to the warm-glow explanation. If, as I suggest below, reward behavior is likely to be weak, third-party punishment might be the best place to look.

8. Some Suggestions

Let me conclude by suggesting some other experiments that might (or might not) be worth while to carry out. As an ignoramus in the field, I cannot tell whether they have already been done, or would be unfeasible, but let me nevertheless put them on the table.

Getting around the 'strategy method' problem. In several studies (e.g. Fehr/Fischbacher 2003; Fehr/Fischbacher 2004b) subjects do not respond to an actual choice by other parties, but to a range of hypothetical choices. As Fehr notes (Fehr/Fischbacher 2004b, 67), a "potential" (and in my view highly probable) disadvantage of this method is that a subject might "experience stronger emotions when reacting to an actual violation of a fairness norm than when contemplating what he would do in case of such a violation". A reason why he nevertheless adopts it is that since, for instance, "dictators rarely or never choose certain transfer levels ... we would have few data for these levels" if only responses to actual choices were possible. This problem could be overcome by having subjects respond to computer-generated transfers at any level, as long as they thought they were dealing with a real person. Given the anonymity of the experiments, this should be easy to achieve. The number of subjects would have to be greater, however, so that financial constraints might count against the idea.

Joining up with prospect theory. In the experiments, punishments have direct costs (out-of-pocket money) for the punisher. Prospect theory asserts that equivalent opportunity costs are perceived as less valuable, i.e. that people are more willing to forgo a gain that to incur a same-sized loss. Since one of the action tendencies triggered by the violation of social norms is ostracism, involving opportunity costs for the ostracizer rather than direct costs, it might be interesting to see whether subjects are willing to incur larger costs of punishments if they take the form of opportunity costs. One might also imagine experiments in which the cost for the punished or ostracized take the form of lost opportunities, e.g. if the punisher has the option of choosing another partner for a subsequent and mutually profitable interaction in which there is no free-rider temptation. The last idea reflects an important fact not captured in Fehr's work, namely that even if defectors are little affected by third-party punishments (Fehr/Fischbacher 2004b, 82), they might be seriously hurt by the unwillingness of third parties to deal with them in the future.

Varying informational conditions. In real-life, the satisfaction from punishing someone has three sources: (i) that a person who harmed oneself suffers a harm, (ii) that the harm is imposed by oneself, and (iii) that one knows the other to be aware of (ii). One might try to determine the importance of the last condition

by seeing whether the willingness to punish is greater when punishers believe that targets of punishment will know that they are being punished. In an experiment suggested above, potential punishers could be told, in one condition, that targets would be told that they were deliberately excluded from the profitable interaction and, in another condition, that targets would be told that they were excluded by a randomizing device. "You can punish him, but he won't know he's being punished." One might conjecture that if punishment occurs in the latter condition, it would have a greater element of impartiality and thus not lend itself as well to the warm-glow explanation.

Introducing positive emotions. Anger, indignation and contempt are negative emotions triggered by wrongdoings. Yet wrongdoers also have *victims*, whose fate might also trigger emotion. In a DG with third-party punishment, for instance, one might give subjects the option of either punishing an ungenerous dictator or transferring money to his victim; they might also be allowed to do both in various proportions. Alternatively, one might give them only the option of compensating the victim and compare patterns with those observed in the punishment-only condition. Also, subjects might be allowed to reward others for supererogatory behavior, either towards themselves or towards third parties. Again, the reward might take the form of being preferentially selected for subsequent profitable interactions. My hunch is that results would align with the legal system, in which wrongdoers are punished, victims compensated, but welldoers not rewarded.

Punishing non-punishers. It is not always clear what counts as a norm-violation, in the operational sense of behavior that would lead to loss of reputation. As Fehr asks, "should an individual who does not help a person with a bad reputation lose his good reputation?" (Fehr/Fischbacher 2003, 789) Along the same lines, should a person who accepts a low offer in a UG be seen as a wrongdoer because he doesn't stand up to the bully? In other words, is there a social norm that under certain circumstances unfairly low offers are to be rejected? Third parties might be given the options of punishing not only those who *make* ungenerous offers in the UG, but also those who *accept* them; alternatively, they might be offered only the latter option. If they choose to avail themselves of it, it would provide evidence that people do punish non-punishers, contrary to what I suggested above.

Bibliography

Bass, G. J. (2000), *Stay the Hand of Vengeance: The Politics of War Crimes Tribunals*, Cambridge

Becker, G. (1957), *The Economics of Discrimination*, Chicago

Coleman, J. (1990), *Foundations of Social Theory,* Cambridge

de Quervain, D. et al. (2004), The Neural Basis of Altruistic Punishment, in: *Science 305*, 1254-1258

Descartes, R. (1985), Passions of the Soul, in: *The Philosophical Writings of Descartes.* Vol. I, Cambridge

Ellickson, R. C. (1991), *Order without Law*, Cambridge

Elster, J. (1989), *The Cement of Society*, Cambridge
— (2004a), Mimicking Impartiality, in: K. Dowding/R. Goodin/C. Pateman (eds.), *Justice and Democracy*, Cambridge, 112–126
— (2004b), *Closing the Books: Transitional Justice in Historical Perspective,* Cambridge
Fehr, E./S. Gächter (2002), Altruistic Punishment in Humans, in: *Nature 415*, 137–140
— /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
— / — (2004a), Social Norms and Human Cooperation, in: *Trends in Cognitive Sciences 8*, 185–190
— / — (2004b), Third-Party Punishment and Social Norms, in: *Evolution and Human Behavior 25*, 63–87
— /B. Rockenbach (2003), Detrimental Effects of Sanctions on Human Altruism, in: *Nature 422*, 37–40
— /J. Henrich (2003), Is Strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Atruism, in: P. Hammerstein (ed.), *Genetic and Cultural Evolution of Cooperation*, Cambridge
Lovejoy, A. O. (1961), *Reflections on Human Nature*, Baltimore
Petersen, R. (2001), *Resistance and Rebellion: Lessons from Eastern Europe*, Cambridge
Proust, M. (1954), *A la recherche du temps perdu*. Vol. III, Paris
Posner, E. (2000), *Law and Social Norms*, Cambridge
Schachter, S./J. Singer (1962), Cognitive, Social and Physiological Determinants of Emotional State, in: *Psychological Review 63*, 379–399