Christopher Stephens

Strong Reciprocity and the Comparative Method

Abstract: Ernst Fehr and his collaborators have argued that traditional explanations of human cooperation cannot account for strong reciprocity. They provide substantial empirical evidence that strong reciprocity is an important phenomenon that cannot be explained by the traditional models of kin selection or reciprocal altruism. In this note, however, I argue that it will be difficult to test specific adaptive explanations of strong reciprocity because it is apparently unique to humans. Consequently, it is difficult to employ the comparative method, which is one of biology's best tools for testing adaptationist claims.

0. Introduction

Human beings display a striking tendency to act in ways that conflict with their material self-interests. They reject unfair offers in one-shot ultimatum games, and they cooperate with strangers in anonymous public goods and prisoner's dilemma games. Fehr and his collaborators (Fehr/Gächter 2002; Fehr/Fischbacher/Gächter 2002; Fehr/Rockenbach 2003; Gintis/Bowles/Boyd/Fehr 2003; Fehr/Fischbacher 2004) have carried out a number of interesting and careful experiments to further support the idea that humans do not always try to act in their own material interests. Their experiments show that humans will cooperate with strangers even in cases where there does not appear to be any reputation benefit from cooperation or hope of future reciprocation. These studies cast doubt on the idea that we can fully explain cooperative behaviour in humans by appealing to traditional models of kin selection, reciprocal altruism, indirect reciprocity, reputation effects, or signaling theory.

These scientists argue, in particular, that we cannot explain a kind of cooperation that they call 'strong reciprocity' by any of these traditional models (Gintis 2000; Fehr/Fischbacher/Gächter 2002). Strong reciprocity is characterized by the tendency to punish defectors and reward other cooperators in public goods games, even at a cost to oneself. The existence of individuals who are willing to punish non-cooperators has two important effects. First, it decreases (or eliminates) the attractiveness of defecting because others will punish you for defecting. The threat of such punishment may even make the cooperative behaviour in the agent's narrow self-interest. Second, the presence of strong reciprocators creates a kind of 'second order' public goods problem. While the existence of humans willing to punish those who don't cooperate may make cooperation pay for every individual, the 'policing' agents are themselves acting

altruistically, because they must pay a cost to punish the defectors. This second order altruism would seem to require some kind of group selection in order to be favoured by natural selection—whether a more narrowly biological model (Sober/Wilson 1998) or a gene-culture model (Richerson/Boyd/Henrich 2003). However, the altruism involved in this second order public goods problem may be significantly less than the altruism in the original first order public goods dilemma. This makes it easier to meet the necessary conditions for group selection to be strong enough to overcome individual selection, which operates against agents that exhibit strong reciprocity.

Fehr and his collaborators' main hypothesis is that strong reciprocity results in (or is partially defined as) the altruistic punishment of defectors and helps encourage cooperation if certain conditions are met. Their experimental evidence suggests that free riders cause negative emotions in others (and that free riders expect this). Feelings of anger toward those who have defected unfairly can be a proximate mechanism that motivates strong reciprocators. One interesting feature of their work is that it is not just any sort of punishment of non-cooperators that will help encourage (make stable) cooperation in the population, it is fair punishment that plays a crucial role in the evolution of strong reciprocity.

We can divide the relevant questions about human cooperation into three categories. First, to what extent do humans cooperate with one another and under what circumstances? Second, what are the proximate mechanisms at work that explain the cooperation? Finally, what is the ultimate, evolutionary explanation of the cooperative behaviour? Fehr and his collaborators have made important contributions to answering all three of these questions. Although I find their main points persuasive, I have some residual concerns about the extent to which their studies bear on these questions, especially the evolutionary one.

2. What Exactly is Strong Reciprocity?

In order to illustrate the difference between strong reciprocity, altruism and reciprocal altruism, Fehr, Fischbacher, and Gächter (2002) use a *sequential one-shot* Prisoner's Dilemma game. In a sequential one-shot PD game, the first player must choose whether to cooperate or defect before the second player, who then cooperates or defects after seeing the first player's choice. According to Fehr, et. al. (2002) an *altruist* is someone who cooperates no matter what, regardless of whether she is the first or second player and regardless of what the other player does. A *reciprocal altruist*, on the other hand, "only cooperates if there are future returns from cooperation" (2002, 3) and so if player two is a reciprocal altruist she will always defect in a sequential one-shot PD. Finally, a *strong reciprocator* will cooperate is she is player one and will also cooperate if she is player two on the condition that player one cooperates.

This way of describing these behavioural types has some peculiar consequences. For example, it can turn out that organisms that play strategies such as *tit-for-tat*—often viewed as reciprocal altruists—are not so on this definition. For one thing, tit-for-tat is usually defined so that it is never the first player to

defect. Tit-for-tat cooperates if the other player cooperates, and so tit-for-tat would be a strong reciprocator rather than a reciprocal altruist in the sequential one-shot PD game.

When theorists such as Trivers (1971), or more recently, Nowak and Sigmund (1993), argue that the success of tit-for-tat like strategies can explain the evolution of reciprocal altruism, they seem to be relying on definitions of strategies that may also be strong reciprocators in different games. In the traditional models, strategies such as tit-for-tat are often favoured in certain games because they tend to encourage cooperation and do well when they play their own kind, consequently thriving in the long run if they interact with their own kind often enough.

The real distinctiveness of strong reciprocity is in third party situations, where the traditional algorithm for tit-for-tat is underspecified. The traditional models tend to focus on two player interactions, and the situations that are most of interest to Fehr and his colleagues are cases where an agent can punish or reward players not only for their interactions with the strong reciprocator, but for their interactions with other players.

Another feature of their definition of strong reciprocity is that it is not clear whether they are talking about *behaviours* or *strategies*. In most games, strategies (such as tit-for-tat) are defined by specifying a behavioural algorithm. But there is a many-many mapping between behaviours and strategies so that more than one strategy can lead to a given behaviour and vice versa. A definition of a reciprocally altruistic strategy might be something like "a reciprocal altruist performs altruistic actions only if *she thinks* the total material returns will exceed the total material costs". This can be contrasted with a more behavioural definition that they give, in which one is a reciprocal altruist only if there is actually a benefit to both players. This difference is revealed in comparing cases where tit-for-tat is, and is not, an evolutionary stable strategy. For example, if all the other players in a standard iterated PD game are defectors, then the tit-for-tat strategy will not be in the agent's long term self interest. So does tit-for-tat count as a reciprocal altruist? If we use a strategic definition, the answer is yes, but on a behavioural definition, the answer is no.

Not much depends on this issue, but I was a little confused by some of their descriptions of reciprocal altruism and strong reciprocity. For example, they say that reciprocal altruists are 'selfish' strategies and that strong reciprocators are 'benevolent'. Humans might be motivated to help others even if such help in fact benefits themselves and vice versa. I think that their talk of motivation should be understood metaphorically here, since they point out in Fehr and Fischbacher (2003) that they don't want to address the motivational debate about psychological egoism. Still, they characterize reciprocal altruists in this way: "Since a reciprocal altruist performs altruistic actions only if the total material returns exceed the total material costs we do not use this term in the rest of the paper. Instead, we use the term 'selfish' for this motivation." (Fehr/Fischbacher/Gächter 2002, 3) But why is this a motivation at all? Why isn't this just a behavioural definition of reciprocal altruism? Indeed, it isn't clear that strategies such as tit-for-tat are 'selfish' since they never do better

than any other strategy with which they are paired (because they are never the first to defect). True, they can do better in the long run, but within each 'group' (where they are paired up with a given opponent for a number of rounds), they are altruists. This is one reason why Sober and Wilson (1998, 81–2) view the success of tit-for-tat in standard reciprocal altruism models as a case of group selection. At any rate, this is probably just a terminological issue, and it is clear enough that the standard models of reciprocal altruism and kin selection are insufficient to handle the phenomenon of altruistic punishment, especially when the group size is large.

3. Motivational Issues

Although perhaps of less interest to economists, many psychologists (Batson 1991) and philosophers (Sober/Wilson 1998) have wondered whether experimental evidence could shed light on the issue of psychological egoism and altruism. The issue here is not whether or why natural selection might favour cooperation, nor is the question one about simply how much humans cooperate with one another. Instead, the issue is whether or not our ultimate psychological motives for our actions are self or other-directed. Although Fehr and his colleagues do not attempt to directly answer this question, one might think that their work bears on this issue. Perhaps the failure of kin selection, reciprocal altruism and signaling theory which they characterize as 'nepotistic' and 'selfish' in motivation, and the necessity of strong reciprocity in explaining certain kinds of cooperation in humans suggests that we can answer the motivational question in favour of psychological altruism (at least in some cases). The debate about psychological egoism is usually cast as one between skeptics who think that there is not sufficient evidence to show that our motivations are ever ultimately 'other directed' in the psychological sense, and those who think that we sometimes are. Everyone agrees that no one is motivated by altruism all the time.

Interestingly, however, the recent work on the neurophysiology by de Quervain, et al. (2004) actually suggests another mechanism (internal rewards) which a defender of psychological egoism could make use of in defending their position. If the brain's reward centre is activated whenever the agent cooperates, then it is possible that this is the ultimate motivation for acting cooperatively in such circumstances. Of course, the mere fact that a pleasure centre in the brain is activated when agents perform altruistic punishment doesn't mean that this is the *reason* why those actions were performed, but it nevertheless provides one more possible mechanism that a defender of psychological egoism could appeal to in defending the view that humans are always ultimately motivated by self-directed considerations. Because these kinds of psychological egoism trace their motivations back to 'internal' factors, it isn't clear whether economists need to worry about this issue, since, if someone is willing to jump on a grenade to save their friends or live like Mother Teresa to help the poor, it doesn't much matter what their ultimate motivation is—if these actions cannot be explained by some traditional material self interest model, then that is enough to force economists

to move to a model where there can be 'social preferences' that don't necessarily involving having more of some material good for oneself.

4. Can We Test the Strong Reciprocity Hypothesis With the Comparative Method?

Defenders of the strong reciprocity hypothesis claim that none of the standard evolutionary models predict that strong reciprocators will be evolutionary stable strategies. Their work, by contrast, predicts that strong reciprocity is plausibly understood as part of a stable polymorphism of reciprocators, cooperators, and defectors. Some (Johnson et al. 2003) have challenged their claim that strong reciprocity cannot be explained by appeal to the traditional models of reciprocity, signaling, or kin selection. They have claimed that humans' current tendency to engage in strong reciprocity might be maladaptive or a side effect of some other trait that was favoured by natural selection.

Fehr and Henrich (2004) reply to this worry at length, invoking both ethnographic information about hunter gatherer societies as well as comparative information about non-human primates. Both sorts of evidence indicate that the maladaptive hypotheses are implausible. One might think that the tendency of some agents to act altruistically in one-shot games is just a side effect of a general strategy of cooperation in case there is some chance one would encounter the same player again. That is, one might think that agents are simply 'hedging their bets' in case there are future opportunities for reciprocity. If most interactions in the relevant evolutionary ancestral environment were generally with the same people, then we could perhaps explain the tendency of people to cooperate in non-iterated public goods games as a mere side effect of a general strategy of assuming that one will interact with the same people repeatedly.

However, Fehr and Henrich (2004) remind us that both contemporary hunter-gatherer societies as well as non-human primates are able to distinguish between strangers and long-term coalition partners. Furthermore, follow up studies on experimental subjects in one-shot games shows that they seem to believe that they will not encounter the other subjects again, so it is not plausible to assume that they simply did not fully appreciate that the game was one-shot.

Fehr and Henrich (2004) also use some comparative information to rule other sorts of non-adaptationist explanations of strong reciprocity. Consider the hypothesis that we simply did not need to know how to discriminate between kin and non-kin in our evolutionary environment and so we mistakenly rely on altruistic principles in our present environment when we are not interacting with kin. Fehr and Henrich (2004) point out that historical information counts against this hypothesis because there is evidence that (1) all primates can distinguish kin from non-kin and consequently reason to think that our evolutionary ancestors could distinguish between them, and (2) we interacted with both in the relevant evolutionary past and it was often important to be able to recognize the difference between kin and non-kin.

The comparative method is important because it can help test crucial his-

torical assumptions that are involved in an adaptive explanation. For example, in explaining why primates and larger mammals generally have larger testes, scientists such as Short (1981) and Harcourt et. al. (1981) suggested that it was a result of sperm competition. In primates where a female has more opportunity to mate with many males, we should expect larger testes. Subsequent studies confirmed the sperm competition hypothesis. In primate species where a given female has the opportunity to mate with many males (as in chimpanzees), the testes are bigger that in species such as gorillas where females have fewer such opportunities. Furthermore, this sperm competition hypothesis makes other predictions that can be independently tested, such as the fact that there should be higher rates of sperm production in species with relatively large testes.

One of the points in favour of the need for an alternative to the traditional models of cooperation is that strong reciprocity is unique to humans (Stevens/Hauser 2004). Animals such as guppies, baboons and vampire bats cooperate, but there is not any evidence that such non-human animals engage in strong reciprocity. Stevens and Hauser (2004) argue that these other animals don't have the cognitive sophistication to engage in this kind of cooperation. This means that traditional explanations that rely, for instance, on standard models of reciprocal altruism won't explain the phenomenon at issue because many non-human animals engage in reciprocal altruism, and the cognitive demands aren't as great. Even guppies appear to play some strategy like tit-for-tat when they go on predator inspection visits (Milinski 1987).

While the fact that strong reciprocity is (apparently) unique to humans raises difficulties for the views that wish to explain all of human cooperation using the same kinds of mechanisms that we use to explain non-human cooperation, it also raises certain difficulties for the adaptationist explanations that Fehr wants to give. This is because it is difficult to use the comparative method to get information about the evolution of a trait that is unique to humans. Notice that in ruling out various *maladaptive* hypotheses, they rely on comparative information to get evidence against crucial historical assumptions that were made by defenders of the 'maladaptive' hypothesis. These maladaptive explanations did not explain cooperation as a side effect of traits unique to humans; rather, their explanations were disconfirmed by the comparative method precisely because they were hypotheses that would help explain human and non-human cooperation.

Since most of our close relatives are extinct, we are unable to engage in the usual procedure of looking deeper into the tree of life and considering traits that we can compare across species that are not all extinct. When it comes to extinct species, information about social structure is harder to come by than information about bone structure. Consider the evolution of sexual dimorphism in humans. Why are human males, on average, bigger than human females? If we embed this problem in a larger context, we can use comparative data to give us more than one data point. For example, Clutton-Brock and Harvey (1977) found a correlation in primate species between the extent to which males are larger than females and the extent to which females outnumber males in breeding populations. This provides support for the adaptationist hypothesis that sexual selection has contributed to sexual dimorphism in humans, as well

as other primates. The worry is that without comparative data of this sort, it is too easy to make up an adaptationist explanation of sexual dimorphism for just one species.

Sometimes we discover that the historical facts show that an otherwise plausible hypothesis is flawed. Some scientists thought that the low birth weight found in bears was a by-product of the evolution of hibernation. An investigation into the phylogeny of bear species, however, revealed that low birth weight evolved before hibernation, and appears on branches of the evolutionary tree on which hibernation never evolved (McKitrick 1993).

The basic idea behind the comparative method is to examine correlations between two or more traits or between a trait and some environmental variable. The extent to which we can test the strong reciprocity hypothesis by the comparative method depends in part on what exactly strong reciprocity is an adaptation for. Here it would be nice to have more details from Fehr and his colleagues. Is the claim analogous to Dunbar's hypothesis (Dunbar 1996) that language evolved to deal with the increased communication demands of living in larger groups? Perhaps strong reciprocity is a result of selection pressures on our ancestors who started living in bigger groups. In order to handle the greater complexity of social interactions, they developed the necessary cognitive resources that then allowed them to engage in strong reciprocity. Or perhaps the greater cognitive resources were selected for in part because of the demands of certain decision problems that required strong reciprocity as a solution. Perhaps there were more situations that demanded cooperation when the entire group was threatened (making it difficult for simple reciprocal altruism to ensure cooperation).

Fehr and his collaborators do give examples of public goods games that were plausibly played in our evolutionary history such as cooperative hunting (though this would not likely be well represented by a one-shot game), food sharing, collective warfare, and the like. The crucial cases to support strong reciprocity are the situations in which there either will not be repeated interactions or there are more than two players acting. In such cases indirect or reciprocal altruism won't be successful in ensuring cooperation. Getting more comparative information would be hard, but it would also help Fehr and his colleagues to discriminate between different adaptive hypotheses about why strong reciprocity evolved.

Here is one final issue to consider. In their public goods games experiments, some players are strong reciprocators and some are not. What explains the variation here? Richerson, Boyd and Henrich (2003) point out that social structures can influence cooperation, so perhaps that is part of the explanation. Wilson (1994) points out, however, that there are at least two possible explanations of such a polymorphism—one is phenotypic plasticity and another is genetic variation. The variation in cooperation might be a result of a kind of phenotypic plasticity that is activated by different rules under different social circumstances, or it might be the case that genetic variation can help explain why some folks cooperate and some do not.

5. Conclusion

A large part of Fehr and his collaborators' work has been to show that strong reciprocity cannot be explained if we assume that the only mechanisms at work are 'selfish' mechanisms such as direct and indirect reciprocal altruism, kin selection, reputation effects, and signaling theory. In order to explain how certain sorts of cooperation became prevalent in humans, we need to consider the role of group selection, especially group selection that is combined with culture as a supplemental means of inheritance of traits between parents and offspring. In this regard I think they have made a convincing case. Strong reciprocity is indeed a phenomenon that we should take seriously. I am more pessimistic, however, about confirming or disconfirming particular adaptive explanations. This is not because I doubt that strong reciprocity is an adaptation; rather, it is because the trait is (among non-extinct species) unique to humans. Because of this, we only have one data point when applying the comparative method.

Bibliography

Batson, C. D. (1991), *The Altruism Question. Toward A Social Psychological Answer*, Hillsdale

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

Clutton-Brock, T./P. Harvey (1977), Primate Ecology and Social Organization, in: *Journal of the Zoological Society of London 183*, 1–39

Dunbar, R. (1996), *Gossip, Grooming, and the Evolution of Language*, Cambridge

Fehr, E./U. Fischbacher/S. Gächter (2002), Strong Reciprocity, Human Cooperation and the Enforcement of Social Norms, in: *Human Nature 13*, 1–25

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

Fehr, E./B. Rockenbach (2003), Detrimental Effects of Sanctions on Human Altruism, in: *Nature 422*, 137–140

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

— /J. Henrich (2004), Is Strong Reciprocity a Maladaptation?, in: P. Hammerstein (ed.), *The Genetic and Cultural Evolution of Cooperation*, Cambridge

Gintis, H. (2000), Strong Reciprocity and Human Sociality, in: *Journal of Theoretical Biology 206*, 169–179

— /S. Bowles/R. Boyd/E. Fehr (2003), Explaining Altruistic Behavior in Humans, in: *Evolution and Human Behavior 24*, 153–172

Harcourt, A. H./P. H. Harvey/S. G. Larson/R. V. Short (1981), Testis Weight, Body Weight and Breeding System in Primates, in: *Nature 293*, 55–57

Johnson, D./P. Stopka/S. Knights (2003), The Puzzle of Human Cooperation, in: *Nature 421*, 911–912

McKitrick, M. (1993), Phylogenetic Constraint in Evolution: Has it Any Explanatory Power?, in: *Annual Review of Ecology and Systematics 24*, 307–330

Milinski, M. (1987), Tit for Tat in Sticklebacks and the Evolution of Cooperation, in: *Nature 325*, 43–35

Nowak, M. A./K. Sigmund (1993), A Strategy of Win-stay, Lose-shift that Outperforms tit-for-tat in the Prisoner's Dilemma Game, in: *Nature 364*, 56–58

Richerson, P./R. Boyd/J. Henrich (2003), Cultural Evolution of Human Cooperation, in: P. Hammerstein, (ed.), *Genetic and Cultural Evolution of Cooperation*, Cambridge

Short, R. V. (1981), Sexual Selection in Man and the Great Apes, in: C. E. Graham (ed.), *Reproductive Bilogy of the Great Apes*, New York

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

Stevens, J./M. Hauser (2004), Why be Nice? Psychological Constraints on the Evolution of Cooperation, in: *TRENDS in Cognitive Sciences 8*, 60–65

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

Wilson, D. S. (1994), Adaptive Genetic Variation and Human Evolutionary Psychology, in: *Ethology and Sociobiology 15*, 219–235