Is Strong Reciprocity a Maladaptation?

On the Evolutionary Foundations of Human Altruism[1]

Ernst Fehr
University of Zürich

Joseph Henrich
Emory University

Forthcoming in: P. Hammerstein (Ed.), The Genetic and Cultural Evolution of Cooperation. MIT Press, Cambridge, Mass.

Abstract: In recent years a large number of experimental studies have documented the existence of strong reciprocity among humans. Strong reciprocity means that people willingly repay gifts and punish the violation of cooperation and fairness norms even in anonymous one-shot encounters with genetically unrelated strangers. We provide historical and experimental evidence suggesting that ultimate theories of kin selection, reciprocal altruism, costly signaling and indirect reciprocity do not provide satisfactory evolutionary explanations of strong reciprocity. The problem of these theories is that they can rationalize strong reciprocity only if it is viewed as maladaptive behavior whereas the evidence suggests that it is an adaptive trait. Thus, we conclude that alternative evolutionary approaches are needed to provide ultimate accounts of strong reciprocity.

[1]This paper is part of a research project on strong reciprocity financed by the Network on Economic Environments and the Evolution of Individual Preferences and Social Norms of the MacArthur Foundation.

In recent years a large body of evidence has emerged from laboratory experiments indicating that a substantial fraction of people willingly repay gifts and punish the violation of cooperation and fairness norms, even in anonymous one-shot encounters with genetically unrelated strangers (see, e. g., Fehr and Gächter 1998a, 1998b; Henrich et al. 2001; McCabe, Rassenti and Smith 1998; Fehr, Fischbacher and Gächter 2002). This behavioral propensity has been termed strong reciprocity (Gintis 2000; Bowles and Gintis 2001; Fehr, Fischbacher, Gächter 2002).

In this paper we discuss the evidence bearing on the question whether strong reciprocity represents adaptive or maladaptive behavior. When the above-mentioned evidence is presented to biologists, zoologists, primate researchers or evolutionary psychologists they often spontaneously provide an ultimate account of strong reciprocity as a maladaptation. They argue that in the evolutionarily relevant past, humans evolved in small groups with frequent repeated interactions and strong reputation mechanisms. In the presence of repeated interactions, or when people's reputation was at stake, they faced a strong fitness incentive to cooperate because in response to their non-cooperative behavior other group members may have refused to engage in profitable future interactions with them, or they may even have punished them directly. Moreover, so the argument goes, because anonymous one-shot interactions have been rare in the past, human psychology tends to 'misfire' in the anonymous one-shot encounters that characterize modern life and laboratory experiments. That is, humans miss-apply behavioral rules, which make adaptive sense in repeated interactions or when their reputation is at stake, to one-shot situations—that is, they 'mistakenly' cooperate and punish in one-shot situations. In this view, strong reciprocity represents a behavioral trait that can only exist in evolutionary disequilibrium. If one-shot encounters become frequent, as it is undoubtedly the case in modern societies, natural selection is expected to reduce the frequency of strongly reciprocal individuals because these individuals will do worse than individuals who do not bear the costs of cooperation and punishment in one-shot encounters.

Despite the superficial plausibility of this maladaptation argument, we believe that there is little evidence in favor of this view and fairly strong evidence against it. We show that there is a lot of laboratory evidence as well as field evidence from small scale societies which contradicts the maladaptation hypothesis. To provide the basis for our discussion we first define strong reciprocity more precisely and present some of the evidence on strong reciprocity that is relevant

for our discussion. Then we take a look at the evolutionary history of humans to see whether it is indeed the case that encounters with no or a low probability of future interactions have been rare.

It is important to stress that this paper deals with the *ultimate* sources of strong reciprocity. Thus, we do not discuss the importance of strong reciprocity for the functioning of friendships, neighborhoods, markets, organizations and the political economy. It has been shown elsewhere that strong reciprocity has decisive effects on the functioning of many aspects of modern societies (Fehr and Fischbacher 2002a; Fehr, Fischbacher and Gächter 2002). This role of strong reciprocity remains true irrespective of whether the maladaptation hypothesis is valid or not.

Our critique of the maladaptation hypothesis does not imply that we consider other ultimate sources of human altruism – kin selection, reciprocal altruism and reputational forces – as unimportant. There is good evidence indicating that they are important. However, our critique means that these other ultimate sources of human altruism do not provide an explanation of strong reciprocity. They provide, therefore, a rather incomplete picture of the evolutionary forces that shaped human cooperation and altruism.

Before plunging into the details, we will briefly describe the typical circumstances under which the experiments that provide much of the evidence for strong reciprocity take place. In these experiments, researchers meticulously ensure that all interactions between the subjects take place anonymously, so that neither before, during, nor after the experiments the subjects were informed about the identities of their interaction partners. There are at least two reasons for the anonymity requirements. First, anonymous interactions provide an interesting baseline case, which, when compared with different types of non-anonymous interactions, allow us to measure the impact on non-anonymity on behavior. Second, if helping or punishing behavior shows up in non-anonymous interactions, one can always argue that because the subjects know each other they might somehow engage in repeated interactions after the experiment. Thus, altruistic helping and punishing behavior cannot be identified in a clean way in non-anonymous interactions. In this large body of experiments, subjects also had to incur real monetary costs when repaying gifts or punishing others, with stakes sometimes approaching three months' income (Cameron 1999; Fehr, Tougareva and Fischbacher 2002). Several of the experiments also ensured that not even the experimenters could observe the individual actions of the subjects (and this was made transparent to subjects). In these experiments the experimenter could only observe the aggregate statistical

results not the behavior of specific individuals. Thus, even when monetary stakes are high, there are no repeated interactions, and subjects can be sure that nobody else knows their behavior (so that reputational factors are ruled out) many subjects exhibited strongly reciprocal responses.

PROXIMATE PATTERNS OF STRONG RECIPROCITY

A person is a strong reciprocator if she is willing (i) to sacrifice resources to bestow benefits on those who have bestowed benefits (= strong positive reciprocity) and (ii) to sacrifice resources to punish those who are not bestowing benefits in accordance with some social norm (= strong negative reciprocity). The essential feature of strong reciprocity is a willingness to sacrifice resources in both rewarding fair behavior and punishing unfair behavior, *even if this is costly and provides neither present nor future economic rewards for the reciprocator*. Whether an action is perceived as fair or unfair depends on the distributional consequences of the action relative to a neutral reference action (Rabin 1993, Falk and Fischbacher 1999).

It is important to distinguish strong reciprocity from 'reciprocal altruism' and from 'indirect reciprocity'. A reciprocally altruistic actor will incur short-run costs to help other individuals only when the actor expects to recoup some long-term net benefits from helping. An indirect reciprocator may also be willing to help if the act of helping can be credibly communicated to others so that the others are more likely to exhibit cooperative behavior towards the indirect reciprocator. Yet, in the absence of repeated interaction or the possibility to gain a favorable reputation these actors never help. This contrasts sharply with a strong reciprocator, who is willing to help another person in response to a kind behavior of this person even in the absence of repeated interactions and opportunities to gain a reputation. The distinction between strong reciprocity, indirect reciprocity and reciprocal altruism can most easily be illustrated in the context of a *sequential* Prisoners' Dilemma (PD) that is played *only once*. Moreover, the game is played under complete anonymity so that any kind of reputation formation can be ruled out. In a sequential PD, player A first decides whether to defect or to cooperate. Then player B observes player A's action after which she decides to defect or to cooperate. To be specific, let the material payoffs for (A, B) be (5, 5) if both cooperate, (2, 2) if both defect, (0, 7) if A cooperates and B defects and (7, 0) if A defects and B cooperates. If player B is a strong reciprocator, she defects if A defected and cooperates if A cooperated because she is willing to sacrifice resources to reward

a behavior that is perceived as fair. A cooperative act by player A, despite the economic incentive to cheat, is a prime example of such fairness. The cooperation of a strong reciprocator is thus *conditional* on the perceived fairness of the other player. In contrast, reciprocal altruists or indirect reciprocators, when in the role of player B, will always defect in an anonymously played sequential *one-shot* PD because in this game there are no future interactions nor is it possible to gain a reputation. This, of course, assumes that players have the ability to comprehend a one-shot game.

The structure of a sequential PD neatly captures the problem of economic and social exchanges under circumstances in which the quality of the goods exchanged is not enforced by third parties like an impartial police and impartial courts. Fehr and Gächter (1998a, 1998b) and Fehr, Fischbacher and Gächter (2002) describe the results of many slightly more general sequential PDs (often called gift exchange experiments or trust experiments) in which the parties are not constrained to pure 'cooperate' or 'defect' choices but can also choose several different intermediate cooperation levels. The upshot of these experiments is that there is a strong positive correlation between the level of cooperation of player A and the level of cooperation of player B. Depending on the details of the parameters, between 40 and 60 percent of the B-players typically respond in a strongly reciprocal manner to the choice of player A: The more A gives/cooperates, the more B gives/cooperates. If player A chooses zero cooperation, then strongly reciprocal B-players also choose zero cooperation. However, there are often also between 40 and 60 percent of second movers who *always* choose zero cooperation irrespective of what player A does. These players thus exhibit purely selfish behavior.

There is an interesting extension of the generalized sequential PD if player A is given the additional option to punish or reward player B after observing the action of player B. In Fehr and Gächter (1998b) player A could invest money to reward or punish player B in this way. Every dollar invested into rewarding increased player B's earnings by 2.5 dollars and every dollar invested into punishment of B, reduced player B's earnings by 2.5 dollars. Since after the reward and punishment stage the game is over, a selfish player A will never reward or sanction in this experiment. In fact, however, many A-players rewarded player B for high cooperation and punished low cooperation. Moreover, subjects in the role of player B expected to be rewarded for high cooperation and punished for low cooperation – consequently, the cooperation rate of player

B was much higher in the presence of a reward and punishment opportunity. Thus, it is not only the case that many B-players exhibit strongly reciprocal responses in the sequential PD, it is also the case that – in the extended version of the sequential PD in which A can punish or reward – the B-players expect the A-players to exhibit strongly reciprocal behavior. This expectation, in turn, causes a large rise in the cooperation of the B-players relative to situation in which the A-players do have no reward and punishment opportunity.

There are many real life examples of the desire to take revenge and to retaliate in response to harmful and unfair acts. One important example is that people frequently break off bargaining with opponents that try to squeeze them. This example can be nicely illustrated by so-called ultimatum bargaining experiments (Güth, Schmittberger and Schwarze 1982; Camerer and Thaler 1995; Roth 1995). In the ultimatum game, two players have to agree on the division of a fixed sum of money. Person A, the Proposer, moves first and makes exactly one proposal of how to divide the amount. Then person B, the Responder, can accept or reject the proposed division. In the case of B's rejection, both players receive nothing, whereas in the case of acceptance the proposal is implemented. In populations from industrialized societies, the result robustly show that proposals that would leave the Responder *positive* shares below 20 percent of the available sum are rejected with a very high probability. This shows that Responders do not behave in a self-interest maximizing manner. In general, the motive indicated for the rejection of positive, yet "low", offers is that subjects view them as unfair. As in the case of strong positive reciprocity, it is worthwhile to mention that strong negative reciprocity is observed in a wide variety of cultures, and that rather high monetary stakes do not change or have only a minor impact on these experimental results. By now there are literally hundreds of studies of one-shot ultimatum games. Rejections of positive offers are observed in Israel, Japan, many European countries, Russia, Indonesia and the US. For an early comparison across countries see Roth, Prasnikar, Okuno-Fujiwara and Zamir (1991). In the study of Cameron (1999) the amount to be divided by the Indonesian subjects represented the income of three months for them. Other studies with relatively high stakes are Hoffman, McCabe and Smith (1996) where US $ 100 had to be divided by US-students, and Slonim and Roth (1998).

Strong reciprocity also plays a decisive role in *n*-person situations that involve the production of a public good. Human history is full of public goods situations, such as cooperative hunting,

food sharing, collective warfare, and the like. An essential characteristic of a public good is that it is difficult, impossible, or not desirable, to exclude those from the consumption of the good that did not contribute to producing it. This then raises the question why anybody should contribute to the provision of the good if the non-providers can also consume the good. In a recent paper Fehr and Gächter (2002) showed that altruistic punishment provides a proximate solution to this problem. A substantial fraction of the subjects in public goods experiments are willing to cooperate *and* to punish the defectors, if given the chance to do so. In these situations, the punishment threat provides an incentive for potential defectors to cooperate, and cooperation flourishes. Whereas, in the absence of targeted punishment opportunities, cooperation typically breaks down, cooperation flourished when targeted punishment of defectors was possible.

HOW PLAUSIBLE ARE MALADAPTATION ACCOUNTS OF STRONG RECIPROCITY?

An important challenge for the maladaptation account of strong reciprocity is the evidence from non-human primates (Boyd and Richerson, in press). Many extant non-human primate species live in small groups very much like those presumed for early humans. In all primate species the members of at least one sex leave their natal groups and join other groups where, in many cases, their only relatives will be their own offspring. It is also well known that primates are able to distinguish kin from non-kin. In most primate groups, there is ample opportunity for repeated interactions among unrelated individuals as well as for reputation formation. However, cooperation among unrelated individuals in primate groups is far less developed than among humans and no behaviors that come close to strong reciprocity have been observed among them (Silk, forthcoming). Therefore, the maladaptation account of strong reciprocity has the problem to explain why repeated interactions among humans lead to strongly reciprocal behavior whereas among primates it doesn't. We conclude from this that the maladaptation account of strong reciprocity is, at least, incomplete.

Furthermore, in zoos and research facilities provisioned primates often live in much larger social groups than they are found in their natural habitats. However, despite being such 'unnatural' social environments, non-human primates do not exhibit any of the 'mistaken' cooperation that is attributed to human living in the larger social groups that characterized modern society. For the same reason that humans mistakenly cooperate in the modern context, the maladaptation hypothesis predicts that non-human primate should 'mistakenly' cooperate in such novel social environments. If non-human primates are not fooled by such unnatural social environments, it seems unlikely that human would be.

The maladaptation account is based on the idea that no ultimate explanation beyond kin selection (Hamilton 1964), reciprocal altruism (Trivers 1971), indirect reciprocity (Alexander 1987, Nowak and Sigmund 1997) or costly signaling (Zahavi and Zahavi 1997; Gintis, Smith and Bowles 2000) is necessary to explain strong reciprocity. In a sense, strong reciprocity is viewed as a byproduct of one of these other ultimate accounts of human cooperation. Proximate mechanisms that have been caused by one of these other ultimate forces are held responsible for the existence of strong reciprocity. These proximate mechanisms, so the idea goes, have been shaped by natural selection but are not sufficiently fine-tuned to the modern human condition where lots of one-shot interactions occur. And they are, in particular, not fine-tuned to the laboratory world of anonymous one-shot experiments. This argument implies that the behavioral rules of humans that produce cooperation and punishment should not be fine-tuned to the distinction between low- and high frequency of future encounters. In other words, humans should exhibit roughly the same behavior in encounters with a high and a low probability of future encounters – and this should be especially true in the conditions of the experimental laboratory. As we will show below this prediction is strongly contradicted by the experimental evidence.

We do not doubt that the other ultimate forces explain important aspects of human cooperation. There is, in fact, persuasive evidence that the nepotistic motives associated with kin selection and the (long-term) selfish motives associated with reciprocal altruism and indirect reciprocity have powerful effects on human cooperation (Silk and Silk 1980; Silk 1986; Daley

and Wilson 1988; Gächter and Falk 2002; Keser and van Winden 2000; Milinski et al. 2002). However, these theories do not provide good ultimate explanations of strong reciprocity. Kin selection could, in principle, account for cooperation in one-shot encounters if humans were driven by rules that do not distinguish between kin or non-kin. Yet, humans like other primates, cognitively and behaviorally distinguish kin from non-kin (Tomasello and Call 1997). Furthermore, people generally feel stronger emotions towards kin than towards non-kin. Parents, for example, have no trouble differentially bestowing benefits on their own offspring, even when their offspring are intermixed with the offspring of others in the same household (Daly and Wilson 1998, Case 2001).

Reciprocal altruism and strong reciprocity

Reciprocal altruism could account for cooperation and punishment in one-shot encounters if the behavioral rules of humans did not depend on the probability of future interactions with potential opponents. However, humans are well capable of distinguishing "partners", with whom they are likely to have many future interactions, from "strangers", with whom future interactions are less likely. There is ample evidence that humans cooperate much more if they expect frequent future interactions than if future interactions are rare or absent (Gächter and Falk 2002, Keser and van Winden 2000, Fehr and Gächter 2000). Gächter and Falk, for instance, conducted sequential PD experiments with many intermediate cooperation possibilities. They implemented a pure one-shot condition in which every player A met ten different B-players (10 different one-shots). In addition, they did a repeated interaction condition in which a pair of A and B-players interacted for 10 periods. The results in the Gächter and Falk study show that the B-players behave much more cooperatively in the repeated condition than in the one-shot condition. Similarly, Keser and van Winden (2000) conducted public goods experiments in a one-shot condition and in a repeated condition. Again, cooperation rates were much higher in the repeated condition. Likewise, Fehr and Gächter (2000) also conducted public goods experiments under one-shot and repeated conditions. Their results also show that cooperation is much higher in the repeated condition –

irrespective of whether there exist opportunities for targeted punishment. For example, when it is possible to directly punish specific other group members subjects contribute 95 percent of their endowment to the public good whereas in the one-shot situation with punishment subjects invest "only" between 60 and 70 percent of their endowment to the public good.

This evidence strongly suggests that laboratory subjects have no problems in understanding the difference between one-shot and repeated interactions. And, the same researchers who spontaneously put forth the 'maladaptation hypothesis' would also likely explain the acuity with which subjects distinguish one-shot from repeated encounters as a result of the 'cognitive architecture' shaped by selective processes favoring reciprocal altruism. In fact, it would be quite surprising if subjects did not understand that the probability of being cheated by a stranger in a foreign town is orders of magnitude bigger than the probability of being cheated by a close friend or a business partner or a colleague at the work place. One of us (Ernst Fehr) has often conducted sequential one-shot PDs in the laboratory. After the experiment, subjects were often disappointed because they failed to exhaust large parts of the potential gains from cooperation. As a consequence, they often complained about the one-shot rules of these experiments saying that it is difficult do establish trust and cooperation with somebody with whom one interacts only once. This all indicates that subjects have no *cognitive* problems in grasping the difference between low- and high-frequency encounters. Yet, perhaps their *emotions* are not fine-tuned to the differences across these two kinds of encounters. It is plausible that emotions like shame or anger enhance the willingness to cooperate and to punish and if these emotions show up regardless of whether we face a one-shot encounter or a repeated encounter they might be responsible for the existence of strongly reciprocal behavior.

Emotions and strong reciprocity

We doubt that our emotions cannot discriminate between, for instance, being cheated by a long-term interaction partner (e.g., a friend) or a short-term interaction partner (e.g. a stranger in a

foreign town). Most of us probably feel much stronger negative emotions if we have been cheated by a friend. To check whether this intuition is correct we conducted a questionnaire among students (n = 172). We asked them whether they felt angrier when a long-term partner had cheated them compared to when a stranger had cheated them. Roughly 80 percent indicated overwhelmingly stronger feelings of anger when the long-term partner cheated them. If anger is indeed the proximate force underlying the punishment of non-cooperators than these answers suggest that the emotional impulse to punish the partner is much stronger.

Sometimes it is also argued that emotions are cognitively impenetrable suggesting that they are overwhelming determinants of human behavior. To support this claim the report of Group 1 (this volume) refers to experiments conducted by Paul Rozin and colleagues. Rozin, Millman and Nemeroff (1986) have performed experiments in which an experimenter gives a subject fudge and then asks the subject (in a between subjects design) if they would be willing to eat more of the same fudge in (a) the shape of a disc or (b) in the shape of feces. Even though the subject knew consciously that the substance is the same fudge they have already eaten (because they could see how the experimenter produced the fudge in the form of discs of feces), most subjects refuse to eat the fudge in the shape of feces. In our view these experiments do not really show that emotions are cognitively impenetrable determinant of human behavior. They only show, that subjects who are nearly indifferent between eating a further piece of fudge or stop eating are affected by the emotion of disgust. We bet if subject were paid for eating the feces-shaped fudge, they would eat the fudge for relatively small amount of money (e.g. $20). More importantly, it seems highly likely that the minimum amount of money (or hunger) which would persuade individuals to eat the feces-shaped fudge would be substantially less than the amount of money that would be necessary to get them to eat real feces in the same shape (we suspect the amount of difference will be infinite). This suggests that the behavioral impulse of emotions is far from being overwhelming or cognitively impenetrable because if the costs of not eating the fudge become sufficiently high, subjects will eat the fudge. Thus, although the existence of emotions

affects our tastes, humans seem to cognitively weigh the costs and benefits of different courses of action, irrespective of whether they are emotion-driven or not.

If this argument is correct and if emotions like guilt, shame and anger are driving forces of strong reciprocity strongly reciprocal behavior patterns should quickly respond to changes in the costs and "benefits". Experiments strongly confirm this argument. Recall the generalized sequential PDs described above in which player A had the option to reward and punish player B after observing player B's choice. Fehr, Gächter and Kirchsteiger (1997) have conducted experiments in which they increased the cost of rewarding and punishing player B by a factor of five. If the A-players emotions are 'penetrable' (capable of understanding costs and benefits in novel situations) as we argue, then they will take into account this cost increase. Therefore, they will reward and punish less. In addition, B-players who understand this, should expect less rewarding and punishment and, consequently, they should cooperate less. Moreover, if the A-players know that – in the high cost condition – any increase in their own cooperation elicits a smaller increase in the cooperation of the B-players, the A-players should also reduce their cooperation. Note, that this argument requires a very subtle chain of reasoning and an understanding of the impact of higher costs not only on one's own (emotion-driven) behavior but also on the (emotion-driven) behavior of the other player. Nevertheless, the behavioral evidence powerfully supports the conclusion that subjects quickly respond to cost changes by adapting their own behavior and anticipating adaptations in the behavior of their opponent: The A-players immediately (before any trial and error learning could occur) punish and reward less in the high cost condition, the B-players immediately expect less punishment and rewarding and, hence, cooperate less. In response to this the A-players also reduce their cooperation significantly.

Another example of instantaneous behavioral changes in response to changes in the benefits is provided by the experiments of Fischbacher, Fong and Fehr (2002). These authors conducted an ultimatum game with competition among responders. Instead of only one responder there were two and five responders. As in the bilateral case the (single) proposer made one offer. Then, the responders simultaneously accepted or rejected the proposal. If all responders rejected, all players

earned zero. If more than one responder accepted the offer, one of the accepting responders was randomly allocated the proposed amount of money, the proposer received the remaining money, and the rejecting responders received nothing. If strong reciprocity were just blind emotion-driven (impenetrable) revenge that is not tailored to the subtleties of the circumstances, one would expect that players in the responder competition condition behave similarly to the bilateral case. If the players' emotions do not understand the difference between one-shot and repeated play why should they understand the much subtler distinction between the bilateral case and the responder competition case? From the viewpoint of the evolutionary history of humans the bilateral one-shot ultimatum game is as artificial as the one-shot game with responder competition. Both games were probably rarely played throughout human evolution and there is thus no reason why human behavior should be well-adapted to the differences across these two games. If, however, there are adaptive reasons for why humans want to punish unfair behavior, that is, if there is an adaptive account for strong reciprocity, then the prediction is different. In the bilateral case the responder basically has a property right in punishment. By rejecting a greedy offer the responder can ensure with certainty that the proposer is punished. The situation is dramatically different in the case of responder competition. Here the responder can no longer unilaterally ensure the punishment of the proposer. In fact, in the two-responder case, a rejection by responder 1 only ensures the punishment of the proposer if responder 2 rejects with certainty, too. Thus, if the adaptive goal of rejections is to punish the proposer we should observe that the responders punish much less in the responder competition condition. Moreover, the reduction in the willingness to punish should be driven by the responders' beliefs about the likelihood that all other responders will punish, too, because only in this case the punishment of the proposer can be ensured. Finally, rational proposers who anticipate that the responders will reject less in the competitive condition will make much greedier offers in this condition. Note that these arguments are again quite subtle and require considerable sophistication.

A great variety of experimental results strongly confirm the strong reciprocity prediction. The responders reject much less in the competitive condition and the proposers respond accordingly.

The rejection rate is highest in the bilateral case, much lower in the two-responder case and even lower in the five-responder case. For instance, whereas in the bilateral case offers of 20 percent of the bargaining cake are rejected with probability 0.8, the same offers are only rejected with probability 0.15 in the five-responder case. As a consequence the average share of the bargaining cake that goes to the responders declines from roughly 40 percent in the bilateral case to 20 percent in the two-responder case, and further to roughly 15 percent in the five-responder case. Moreover, the decline in the rejection rate across conditions is *exclusively* driven by the responders' beliefs about the other responders' rejection behavior. If the responders in the competitive condition believe that all other responders also punish a greedy offer, the probability of rejection in the competitive case is as high as in the bilateral case. Yet, if the responders believe that some of the other responders accept a greedy offer, their willingness to reject the offer becomes much lower – as predicted by the strong reciprocity approach. Thus, it is not the case that because of the competitive situation the responders *somehow* exhibited more competitive preferences, which induced them to reject less often. Instead, the reduction in the willingness to reject is exclusively due to the change in beliefs – as predicted by the strong reciprocity approach.

Reputation formation and strong reciprocity

Indirect reciprocity can only account for cooperation in one-shot encounters if our behavioral rules were not contingent on the likelihood that our actions will be observed by others. Again, there is strong evidence to the contrary. The experiments of Milinski et al. (2002) show, for instance, that cooperation in public goods games breaks down if the possibility to gain a favorable individual reputation is removed whereas if subjects can gain individual reputations, cooperation flourishes. People are well-calibrated to observing opportunities for reputation formation, even in novel laboratory environments – they aren't 'fooled' into thinking reputation is always important, as the maladaptation hypothesis would propose. This break down of cooperation is, by the way, regularly observed in repeated anonymous public goods experiments where reputation formation and punishment is ruled out. This break down is observed even if the same group of people can

stay together for the whole experiment. Thus, in this context the problem is not to explain why humans cooperate but why humans cannot maintain cooperation. It can be shown that the peculiar patterns of strong positive reciprocity provide a plausible explanation of the break down of cooperation.

The peculiarities of strong positive reciprocity in public goods games have been examined by Fischbacher, Gächter, and Fehr (2001). In their experiment a self-interested subject is predicted to defect fully irrespective of how much the other group members contribute to the public good. However, only a minority of subjects behave in this way. About 50 percent of the subjects are willing to contribute to the public good if the other group members contribute as well. Moreover, these subjects contribute more to the public good when others are expected to increase their contributions – indicating a strongly reciprocal cooperation pattern. However, only 10 percent of these subjects are willing to match the average contribution of the other group members whereas 40 percent of the strongly reciprocal types contribute less than the average contribution of the other group members. Roughly 30 percent of the subjects behave in a fully selfish manner – defecting always irrespective of how much they expect others to contribute. The remaining 20 exhibit other patterns.

Based on the patterns of reciprocal behavior observed in Fischbacher, Gächter and Fehr (2001), the break down of cooperation in repeated public goods experiments can be neatly explained by the dynamics of the interaction between strongly reciprocal strategies and selfish strategies. For any given *expected* average contribution of the other group members in period *t*, strong reciprocators either match this average contribution or contribute somewhat less than the expected average contribution. Moreover, the selfish types contribute nothing. Thus, the *actual* average contribution in period *t* clearly falls short of the average contribution that has been expected for period t, inducing the subjects to reduce their expectations about the other members' contributions in period t+1. Yet, due to the presence of reciprocal types, the lower expected average contributions in period t+1 cause a further decrease in the actual contributions in t+1. This process repeats itself over time until very low contribution levels are reached.

It also has been shown that the punishment behavior of laboratory subjects is contingent on the possibility of building a reputation. Fehr and Fischbacher (2002b) conducted a series of ten ultimatum games in two different conditions. In both conditions subjects played against a different opponent in each of ten periods. In each period of the baseline condition, Proposers knew nothing about the past behavior of their current Responders. Thus, the Responders could not build up a reputation for being "tough" in this condition. In contrast, in the reputation condition, Proposers knew the full history of behavior of their current Responder, i.e., the Responders could build up a reputation for being "tough". In the reputation condition, a reputation for rejecting low offers is, of course, valuable because it increases the likelihood to receive high offers from the Proposers in future periods.

If the Responders cognitively and emotionally understand that there is a pecuniary payoff from rejecting low offers in the reputation condition, one should observe higher acceptance thresholds in this condition. If, in contrast, subjects do not understand the logic of reputation formation and apply the same habits or cognitive heuristics to both conditions one should observe no systematic differences in Responder behavior across conditions. Since the subjects participated in both conditions it was possible to observe behavioral changes at the individual level. It turns out that the vast majority (82 percent) of the Responders increase their acceptance thresholds in the reputation condition relative to the baseline condition. Moreover, the increase in acceptance thresholds occurs *immediately* after the reputation condition is introduced. There is not a single subject that reduces the acceptance thresholds in the reputation condition relative to the baseline in a statistically significant way. This again contradicts the hypothesis that the subjects do not understand the difference between anonymous and non-anonymous play. Subjects seem keenly attuned to the differences in reputation building opportunities.

It is sometimes argued (e.g., during the discussions with members of Group 1 of the Dahlem conference) that the application of different behavioral rules across one-shot and repeated interactions, and across anonymous and non-anonymous interactions, do not refute the maladaptation account. Perhaps subjects have a kind of baseline belief such that when they are put in an anonymous one-shot experiment they actually believe with *positive* probability that their actions will not remain anonymous or that they do in fact play a repeated game. Then, if one changes from the anonymous one-shot situation to a non-anonymous situation or to a situation

with repeated play their assessment of the probability of repeated interaction go up. This change in beliefs could then be responsible for the different behaviors. We have tested this argument by asking subjects after each session of a series of anonymous one-shot public goods experiments whether they believed the experimenters' claims in the instructions regarding anonymity and one-shot play. We had the following introductory paragraph for our questions: "Unfortunately, it happens from time to time that our experimental instructions are not sufficiently precise, or that participants forget or do not believe certain aspects of the instructions. To improve our instructions for future experiments we ask you to answer the following questions". After this introductory statement there was the following statement regarding one-shot play: "In the instructions we told you that in every period of the experiment you will be matched with three new persons. However, we do not know whether – during the experiment – you also perceived this in this way". Then the subjects had to indicate YES or NO for the following statement: "During the experiment I assumed that in every period I will be matched with three new persons". After subjects had answered the first question we had the following statement regarding anonymity: "In the instructions we told you that the other group members will never be informed that you have been together with them in a group, that is, the other group members do not receive information about your personal identity. We do not know whether – during the experiment – you believed this". Then they had to indicate YES or NO to the following statement: "I assumed that the other participants will not receive any information about my personal identity."

Ninety-six percent of the subjects (n = 120) indicated that they believed that they will be matched with new persons in every period and that their anonymity will be preserved. This suggests that carefully written instructions in combination with the experimenters' credibility to always say the truth (which is unfortunately often not the case in psychology experiments) induce subjects to believe the anonymous one-shot character of experiments. Yet, sometimes it is argued that the belief that there is a positive probability that one's identity will be revealed or that one plays in fact a repeated game may be driven by unconscious mechanisms. It is, however, often not clear what these unconscious mechanisms could be and how they work. In the absence of more concrete specifications this argument remains elusive and it is difficult to refute because if one provides evidence challenging argument A it is always possible to say, "oh, I didn't mean

argument A but argument B."[2] In fact, to our knowledge, the maladaptation account has never been carefully formulated in terms of empirically refutable predictions. According to our experience, it often comes in the form of vague intuitions based on imprecise and questionable generalizations about human evolutionary history.

If indeed unconscious mechanisms are the reason for helping and punishing responses, why do subjects respond *so quickly* to changes in the cost of helping or punishing? Likewise, why do subjects *instantaneously* change their behavior when repeated interactions or the possibility of reputation formation is introduced. These quick behavioral changes are almost certainly mediated by sophisticated evaluations of the costs and benefits of different courses of action that are available to them.[3] We find it hard to reconcile subjects' quick responses to treatment changes, which almost surely are mediated by sophisticated, conscious, cognitive acts with the view that a cognitively inaccessible mechanism drives the baseline pattern of reciprocal responses.

Another problem with the above-described 'greater than zero baseline probability of repeated interaction' argument is that it can only arise from a fundamental misunderstanding of evolutionary models of repeated interaction. The canonical models of the evolution of cooperation via repeated interactions specify that the selection of a cooperative strategy depends on the probability of future interaction and the fitness costs and benefits. If the ratio of costs to benefits is sufficient large, no amount of repeated interactions (give finite lifetimes) will favor reciprocating strategies. Given that life in ancestral environments provides a full range of costs and benefits of different kinds (from giving one's life for a non-relative in warding off a predator to providing an ounce of meat from a large kill), evolutionary psychology should predict that individuals are able to switch from a zero baseline (full defection) to full cooperation, depending on the detail. Similarly, if groups are fluid and migration is common, the expected number of future interaction will vary on a case by case basis, so individuals should be geared-up to completely withdraw substantial forms of cooperation (high cost/benefit ratio) from ephemeral

[2] If emotions are thought to be the mechanism the claim is doubtful because emotions like say, gratefulness, which enhance cooperative responses to cooperative acts, or anger, which provides a basis for strong negative reciprocity, are not cognitively impenetrable.

[3] For example, when one introduces a cost change within an experimental session subjects have to consciously understand what, say, higher costs of punishment mean for their earnings. Subjects in the experiments are always paid at the end of the whole session. Thus, they do not directly experience these higher costs during the session because the gains and losses accruing during the session are just documented on their decision sheets or in a computer file. Only at the end they experience the higher costs in terms of lower cash earnings.

individuals. Again, evolutionary psychology should predict that individuals are fully capable of grasping a zero baseline.

The maladaptation account also has the problem to explain the fundamental heterogeneity in subjects' behavior. In most experiments there is a large group of subjects indicating strongly reciprocal responses *and* a large group behaving in a completely selfish manner. The relative size of these groups depends on the economic costs and benefits of strongly reciprocal actions. In fact, it is almost surely possible to wipe out any reciprocal responses by making them sufficiently expensive. Yet, if the costs are not too high many subjects reciprocate. How can the maladaptation account explain the existence of completely self-interested behavior? If the maladaptation account is correct why do we not observe everybody engaging in strongly reciprocal behavior?

The frequency of "one-shot" interactions

The laboratory evidence as well as casual observations from everyday life suggest that humans have fine-tuned behavioral repertoires taking into account whether they face kin or non-kin, partners or strangers, and whether they can or cannot gain an individual reputation. This suggests that humans who exhibited these behavioral traits had an evolutionary advantage over those humans who exhibited more blunt behaviors. The likely reason for this advantage is, contrary to common mythology, that humans faced many interactions where the probability of future interactions was sufficiently low to make defection worthwhile. In addition, the costs of mistakenly treating unrelated individuals as kin, or treating strangers as partners, were very high. That is, if one were to "reverse engineer" from the available empirical evidence back to the ancestral environment (the EEA—Environment of Evolutionary Adaptiveness) that characterized human evolution, one would predict that our ancestors had frequent encounters with strangers, and that these encounters had substantial fitness consequences. As we'll illustrate with ethnographic data from small-scale societies, a lack of vigilance in interactions with unfamiliar individuals often had deadly consequences. Using data from both the primatological and ethnographic records, we find no support for the idea that the EEA lacked fitness-relevant interactions with strangers (by 'strangers' we mean individuals who were neither kin nor long-term interactants). To the contrary, data from small-scale foraging societies and chimpanzees

clearly shows that interactions with strangers were likely common and highly fitness relevant.[4] Yet, before we sketch this evidence it is important to discard the false dichotomy between completely anonymous one-shot interactions on the one side and non-anonymous repeated interactions on the other side.

The real question is not whether 100 percent non-repeated interactions (completely anonymous) have been frequent in evolutionary history or not – those unfamiliar with the mathematical details of the relevant theory often couch the argument in these terms. Instead, the really important distinction is between encounters with a sufficiently low probability of future encounters, such that defection was the fitness-maximizing strategy, and encounters with a sufficiently high probability of future encounters, such that cooperation was the fitness-maximizing strategy. If early humans faced a mix of situations such that sometimes the best strategy was defection and sometimes the best strategy was cooperation, then there is an a priori case for the selection of strategies that do *distinguish* between these two situations (this also implies that individuals should be fully capable of taking a 'zero-baseline' of cooperation). Under these circumstances, the logic of the maladaptation argument implies that selection should have favored strategies that can distinguish encounters that have a sufficiently low probability of repeated interaction, and generate non-cooperating, non-punishing behavior in one-shot experiments. Therefore, in these circumstances, the maladaptation argument cannot account for strongly reciprocal responses in one-shot encounters.

To be precise, assume that if the probability of future encounters in a simultaneously played PD is below $p = 0.7$, the fitness-maximizing strategy is to defect, whereas if the probability is above 0.7, the fitness-maximizing strategy is to cooperate.[5] Then individuals who defect in the first and cooperate in the second situation have a selection advantage over those individuals that cooperate in both situations. Thus, selection should favor the individuals who discriminate

[4] Evolutionary psychologists have been notoriously unclear about the statistical reliable patterns that make up the EEA, and more importantly, they have been unclear about what data supports these supposed patterns.

[5] We abstract here from the multiple equilibrium problem that is inherent in any repeated PD or public goods game with a sufficiently high continuation probability. The maladaptation account typically forgets that repeated interactions are by no means sufficient for cooperation, even if the probability of future encounters is one, because there are typically infinitely many equilibria implying less than full cooperation. Thus, even if it were true that humans never faced encounters with a low probability of future interactions, it is not guaranteed that evolution favors strongly reciprocal behavior patterns. The reason is simply that there are also many equilibria in repeated games with non-cooperative outcomes.

between the two situations. Yet, if an individual is capable of understanding that defection is the fitness-maximizing strategy in case of a probability smaller than $p = 0.7$, why should this individual then be unable to understand that defection is also the best thing when the continuation probability is zero. In fact, if the continuation probability is zero it is transparent that there are not future gains from cooperation while if the continuation probability is, say 0.5 the situations is more ambiguous and the costs and benefits from different actions are more difficult to assess. The upshot of this argument is that we believe that interior continuation probabilities have been common in evolutionary history and that evolution has, therefore, favored discriminating strategies. Yet, if humans have been selected to apply discriminating strategies in the case of interior probabilities, that is, they defect even when the continuation probability is positive but insufficiently high, then the discriminating strategy also induces them to defect in the case of a zero continuation probability. Therefore, the maladaptation view cannot account for the existence of strongly reciprocal behavior.

How plausible is the assumption that interior continuation probabilities have been common in our evolutionary history? To get a first grip on this question consider again a simultaneously played PD in which players A and B receive the payoff (c, c) if both cooperate, (d, d) if both defect, (c+t, c–s) if A defects and B cooperates and vice versa in the opposite case. To ensure that the conditions of the PD hold the payoffs obey $c+t > c > d > c–s$. t measures the gains of the defector arising from a deviation from joint cooperation and s measures the loss of the cooperator if the other player defects. It can be shown that in a cooperative equilibrium obtained by so-called trigger strategies the continuation probability has to exceed a critical level p_o which is given by $p_o = t/(c – d + t)$.[6] Thus, the higher the temptation t, the higher is the critical threshold p_o, and the

[6] Consider a cooperative equilibrium in which the joint cooperation outcome is supported by so-called trigger strategies. In a trigger strategy equilibrium both players start with cooperation. They also cooperate in period $t > 1$ as long as nobody has defected in the past but in case that one of the players has defected in the past both defect in period t. Thus, defection by the other player is punished with future non-cooperation but the defector also does not cooperate in the future because he expects the cheated player to defect in the future. A trigger strategy is thus unforgiving and imposes large costs on the defector implying that even for a relatively low continuation probability punishment can be sufficiently high to ensure cooperation. Thus, our assumption of a trigger strategy equilibrium favors the maladaptation view because we allow for large punishments, which render cooperation already an equilibrium at relatively low continuation probabilities. The lower the continuation probability that is necessary to sustain a cooperative equilibrium the more likely human cooperation is achieved in repeated interactions, and the more likely the maladaptation argument applies. How large is the threshold continuation probability that ensures the existence of a cooperative equilibrium sustained by trigger strategies? To compute this probability we first derive the expected payoff – viewed from period τ onwards – of an individual in the joint cooperation equilibrium. With probability p the individual receives in each future period c implying that the expected payoff is given by the

higher the increase in gains arising from mutual cooperation (c-d) the lower is the critical threshold. If the temptation becomes very large and approaches plus infinity the critical threshold approaches 1, rendering cooperation among selfish individuals highly unlikely. In times where the survival of an individual is at stake the temptation value can be plausibly set equal to "plus infinity". Note also that if c-d becomes very small the critical threshold again approaches $p_o = 1$.

All the characteristics of repeated encounters that affect c–d and t affect the critical threshold. It is plausible that throughout evolutionary history humans faced a wide variety of different conditions shifting the threshold up and down but the threshold was always in the interior because $t > 0$ prevailed. Therefore, for sufficiently low continuation probabilities, i.e., if $p < p_o$, defection was the fitness-maximizing strategy. Moreover, due to variation in conditions over individuals' life-times and across difference contexts, ancestral humans probably faced many situations in which defection was worthwhile, even when the probability of future encounters was relatively high – e.g., during times of scarcity, families could offer their children as food to be cooked and consumed by unrelated individuals who would be expected to offer their kids in the next scarcity. Surely, selection would favor a 'defect' strategy in this context, as the benefit/cost ratio is too low.

To assemble an understanding of the relevant characteristics of the EEA, we marshal evidence from both contemporary foraging populations (and other small-scale societies) and extant non-human primates to show that ephemeral contacts with strangers (low repeat interactants) were likely both frequent and carried substantial fitness consequences. We think that this evidence should lead evolutionarily minded scholars to predict that humans should be equipped with specialized cognitive machinery capable of distinguishing low frequency interactants (when $p < p_o$) from long-term repeat interactants. Above, we summarized a substantial body of empirical evidence showing both people do seem fully capable of adapting their behavior in response to information about the likelihood of future interaction, even when that information comes in the form of 'novel' laboratory cues. Thus, we think that much of the laboratory-observed behavior results from adaptive processes acting on human psychology over

---

discounted sum $\Sigma p^{\tau} c$, which is equal to $c/(1–p)$. If, instead, an individual defects in period τ, it receives c+t in this period, but from the next period onwards the payoff is only d. Thus, in case of defection in period τ the expected payoff from τ onwards is given by $(c+t) + p\Sigma p^{\tau} d$ which is equal to $(c+t) + (pd)/(1–p)$. If the expected payoff from

the course of hominid evolution. In what follows, we sketch the basic pattern of evidence, but since the facts are spread over a large number of ethnographies and primatological sources we do not provide a full picture here.

Hunter-gatherers vary widely in the settlement patterns, social structure, economic integration, institutional forms, population density and resource use (Kelly 1995; Arnold 1996). For our purposes, we focus the on nomadic and semi-nomadic foraging populations that are *assumed* to give the most insight into Palaeolithic lifestyles.[7] Ethno-linguistic groups of hunter-gatherers typically number at least in the low thousands, but populations between 10,000 and 15,000 are not uncommon. At the local level, many nomadic foragers live in mobile bands of approximately 25 individuals. The membership of these bands is fluid, and constantly shifting as individuals move from one band to another for various reasons. These local groups often aggregate around centralized resources during certain seasons (e.g., water holes, pinon nuts, annual herds etc.). In these recurrent circumstances, social life becomes very intense: mates are found, ancestors glorified, alliances formed and scores settled. Once every twelve years or so, Shoshone families would aggregate in larger groups for "antelope drives", consisting of 75 people or more (Johnson & Earle 2000).

More irregularly, periodic environmental fluctuations (floods, windstorms, plagues, droughts, hurricanes, volcanic eruptions) often (in evolutionary terms) bring together substantial number of strangers during fitness-critical times, as individuals and groups travel over great distances in search of water, caribou and other resources. In the Kalahari, for example, the most severe droughts hit on average about two per hundred years – meaning the average !Kung experienced one in his lifetime (Lee 1998: 79-83). Under these conditions many of the "permanent water holes" failed, and bands had to travel great distances. Of the five permanent waterholes in the 64,000km^2 Dobe area, only two have never failed in living memory. During these kinds of

joint cooperation, c/(1–p), is higher than from defection, (c+t) + (pd)/(1–p), it is rational for a selfish individual to cooperate. Simple manipulations of this inequality show that this is the case if p > [t/(c – d + t)].

[7] This assumption, however, deserves substantially more scutiny than it typically receives. Most of the foragers used as „models" of EEA societies live in the remote marginal environments leftover after 10,000 years of the agricultural revolution. The few historical and archaeological cases we have of hunter-gathers living in rich non-marginal environments show societies are substantially denser, larger, and more complex than any of the ‚standard models'. The marginal environments inhabited by these remnant foragers are those few places on Earth that resisted plant domestication. The inhabitants of these marginal environments, the smallest-scale, nomadic foragers know, provide the „worst cases" for our argument. If we used the full spectrum of foraging societies, with their high densities, slaves, classes and chiefs, the ‚maladaptation hypothesis' would do even worse.

aggregations, people will encounter many strangers that they are not likely to encounter again in their lifetime. This was actually observed by J. Marshall during the severe drought of 1952, when seven San groups converged on a water hole that many of them hadn't used in living memory, and were unlikely to use again in their lifetimes (Lee 1998:86). Obviously, these events had substantial fitness consequences as large aggregations of strangers attempted to share water, game and monogongo nuts. Weissner (1982) described how the combination of high winds (which destroyed the mongongo nut crop) and a plague (which decimated the meager domestic stocks) caused the local population to scatter itself across numerous camps all over a area. Similar drought-related patterns are around throughout Australian and American ethnographic and historical record (e.g., Cane 1990: 157; Aschmann 1957:96; Peterson 1975, 1978; Peterson and Long 1986). If nothing else, environmental shocks would have guaranteed that strangers encountered one another during fitness critical times, and had to divide resources (or fight) in some fashion. The reader should keep in mind that environments were likely substantially more variable in the Palaeolithic than the Holocene (now), so such population mixes driven by shocks would have likely been more common.

The !Kung institution of *hxaro* (dyadic, long-term trading partnerships) provided a risk-managing means of dealing with both spatial and temporal variation in food and water. In times of acute stress, individuals in families often 'activated' a trading partnership by travelling 200km to visit one of their trading partners, and stay for several weeks. The diffuse networks of *hxaro* trading relationships combined with both temporal and spatial variation, guaranteed a well-mixed population, which means that individuals were likely to encounter strangers in the camps of their hxaro partners. It's often forgotten that while an individual had direct, long-term, reciprocal ties with the *hxaro* partner, they did not have such ties with all the other members of their partners' encampments, with whom they, nevertheless, shared water, game, meat and other local resources as if they were a member of the group. Further, in managing the hxaro relationship, of which the average adult with mature children had around 24 (that's 48 per couple), the average !Kung had to travel an area of at least 10,000 km,^{2} wherein they could encounter easily more than thousand people. However, 10% of hxaro partnerships (i.e., 2.4 of each person's 24 partners) ranged over roughly 140,000km^2, where they could meet up to 14,000 inhabitants. It's hard to believe that

these individuals did not have many cooperation opportunities that were associated with low frequencies of future interactions.

Again, using Kalahari data, Harpending calculated the average distance between the birthplaces of spouses and the birthplaces of parents and offspring. The average distance between parent and offspring is 66.5 kilometers, and over 10% of the population had distances between 100km and 220km. The average distance between the birthplaces of mates is 70km (Harpending 1998). This means that in searching for a mate, the average individual covers a territory of 15,000 km^2, an area in which he or she was likely to encounter 1,500 people. It's essentially impossible that any individual maintained long-term repeated interaction with so many people.

Foragers, sometimes alone or in small groups, are known to have travelled extensively over large regions for a variety of reasons, often to obtain specific resources such as particular kind of wood, or ochre. Australia is perhaps the best place on Earth to look for evidence on this matter, because, until European conquest in the late 19^{th} and early 20^{th} century, much of Australia was inhabited entirely by full-time foragers (Agriculture first arrived with Europeans in Australia). In the Western Australian Desert, particular areas attracted a constant influx of thousands of individuals, who had travelled 100 of miles to obtain a particular type of acacia wood used in crafting fine spears. Myers (1986) has recorded life histories from dozens of Pintupi Australians that had not seen a white man until their later years. These life histories tell of substantial travelling and constant encounters with complete strangers. Sometimes these encounters were friendly and an exchange of goods or information occurred, while at other times hostilities ensued. There seems little doubt that many of these were one-shot interactions with strangers in the middle of the desert, hundreds of miles from one's home territory.

Further, diverse hunter-gatherer groups have rituals for bring strangers into the camp. In several cases, these rituals were observed, or experienced, by the earliest European explorers (and recorded in detail by later ethnographers). On the continent of hunter-gatherers (Australia), ethnographers and explorers found elaborate rituals that were *specifically* used for bringing *strangers* peacefully into camp (Thomson 1932: 163-64). Strangers not performing these rituals were treated as having hostile intentions and usually killed (note the fitness effects, the ability to 'treat' strangers differently). If foraging life doesn't involve encountering strangers, why would nearly all aboriginal groups have elaborate culturally-evolved rituals specific for admitting

strangers? Moreover, this basic ritualized interaction managed to diffuse over a vast region, so it's hard to believe this was a newly developed practice. On the other side of the world, inhabiting the fierce climate on the southern tip of South American (Tierra del Fuego), two foraging groups – the Ona and Yaghan – maintain a similar (though less elaborated) ritual process for bringing strangers in camp. At first European contact, several of Magellan's crew, in search of provision, approached a small encampment of these foragers during their trip through the Strait of Magellan. Unfortunately, these men did not perform the proper ritual and were immediately killed. These lonely foragers immediately spotted these strangers, and reduced their fitness substantially.

Far to the north, arctic foragers travelled extensively in order to maintain knowledge of geography and ecology over vast regions. The Nunamiut maintained knowledge of nearly 250,000 square kilometers (Binford 1983), while typically only using about 25,000 km^2. During their occasional recognizance travels they may have encountered any of the 5,000 inhabitants – and it seems unlikely that each hunter knew and personally maintained long-term relationship with these thousands of adults.

Another line of evidence comes from archaeological and ethno-historical evidence of warfare and long-distance raiding. As we showed for Australia, foragers often travelled long distances to trade or obtain resources from distant groups of strangers, and they encountered many strangers along the way. Human groups, including foragers, have for as long as we can see back in the archaeological record had violent interactions will other groups (Keeley 1996). As far back as the Upper Palaeolithic, we see evidence of large settlements behind defensive walls (Johnson & Earle 2000). For our case, the most persuasive evidence comes from long-distance raiding. Here war parties travelled hundred of miles to raid, pillage, and steal wives from strangers. North American ethnohistorical data show that groups like the Tlingit, from the Alaskan panhandle, raided as far south as Puget Sound, and the Mohave raided groups on the Californian coast. During the historic period, the Iroquois raided Delaware, the Great Lakes and the Mississippi Valley. This kind of raiding, if present in the EEA (there's no reason to think it was not), would have provided a consistent selection pressure to distinguish strangers (possible raiders) from friendly locals. The idea that human psychology should have evolved to cause individuals 'hedge their bets' by assuming that at any particular stranger might be a long-term interactant seems mistaken. Combining this raiding data with the above evidence suggests that people may well have

routinely encountered both strangers who are friendly and interested in trading, and stranger who were dangerous and bent on raiding, stealing and murder. What we don't see in the ethnographic data is any hint that people never encountered strangers in fitness relevant circumstances.

Non-human primates provide another other line of evidence that evolutionary psychologists use (or should use) to develop an understanding of the EEA. In this case, the evidence is stark. The mating patterns of non-human primates means that at least one sex leaves their natal group to find and join another social group. Upon arrival at their new group (which may or may not contain any kin), the existing group members do not confuse the individual with their kin, or coalition partners. The individual has to build her own coalitions gradually, find a mate or mates, and produce some kin. In non-human primate societies, even in the unnaturally large and provisioned ones in zoos and research centers, animals do not 'get confused' and mistakenly treat strangers as long-term partners and kin.

Chimpanzee, as our closest genetic relatives, provide any instructive example. Chimpanzee social groups are known to go on 'night patrols' along the border of their territories. If they encounter a smaller group of 'stranger chimpanzees' (from another chimpanzee group), they attack and kill (or drive off) the strangers (Manson & Wrangham 1991).[8] Chimpanzees don't get confused and treat strangers like long-term coalition partners, other group members, or kin. Chimpanzees apparently have the ability to distinguish group members from strangers, and treat them differentially. Chimpanzee don't have the "non-zero baseline" of prosociality towards other chimpanzees, despite their long history in small-scale societies based on reciprocity and kinship. From this perspective, it seems odd that evolutionary psychologists would argue that humans carry around a non-zero baseline of prosociality for strangers, while chimpanzees do not.

While the above account necessarily lacks the kind of systematic rigor and evidence with which one would like to address the EEA question (because such evidence simply doesn't exist), we believe that it, nevertheless, demonstrates that the widely held view of ancestral human societies as isolated groups that did not mix or interact with surrounding social groups or strangers, has little, if not no, empirical support.

[8] Female strangers may have been incorporated into the raider's group.

CAN STRONG RECIPROCITY SURVIVE IN EVOLUTIONAR EQUILIBRIUM?

Bad theories often survive if no other theories are available that can replace them. Therefore, if strong reciprocity is unlikely to be a maladaptive trait, it is important to develop an adaptive account of strong reciprocity. Recently Price, Cosmides and Tooby (2002) argued that the punishment of non-cooperators in public goods situations evolved because the punishers can reduce or overturn the payoff differences between themselves and the punished defectors. Unfortunately, this argument is theoretically and empirically invalid. It is theoretically invalid because it does not solve the core problem why non-punishing cooperators do not replace the punishing cooperators. This question is decisive because non-punishing cooperators will reap the benefits created by the presence of punishing cooperators, without paying the cost. On the empirical side, experiments conducted by Falk, Fehr and Fischbacher (2001) show that if the punisher cannot reduce the payoff differences – because every dollar invested into punishment only reduces the payoff of the punished by one dollar – the willingness to incur costs to punish defectors is still very high. Thus, a lot of punishment occurs even if payoff differences cannot be affected.

On the positive side, several theoretically rigorous models have been proposed that provide an adaptive evolutionary foundation for strong reciprocity. Taking advantage of the fact that humans, unlike other primates, are heavily dependent on certain kind of imitation and other forms of social learning, Henrich and Boyd (2001) show that an arbitrarily small amount of conformist transmission makes cooperate-punish a stable culturally-evolved equilibrium in one-shot N-person interactions. They go on to show that, once this equilibrium spreads (a processes modeled in Boyd & Richerson 2002), within-group individual selection will favor prosocial genes that allow individuals to avoid the costs of being punished—these are genes that would otherwise not be favored without the interaction of genes and culture (also see Henrich et. al this volume and Richerson et. al. this volume; Henrich forthcoming).

More recently, Boyd, Gintis, Bowles and Richerson (2002) used a simulation to show that a model of cultural group selection is able to explain altruistic cooperation as well as altruistic punishment in large groups. They calibrate the parameters of their model to mimic the likely evolutionary conditions of humans. The idea behind their model is that in the vicinity of an evolutionary equilibrium, where altruistic cooperators and altruistic punishers are frequent,

within-group selection against altruistic punishers is very weak because non-cooperation rarely occurs and, hence, little punishment costs have to be born by the altruistic punishers (a logic first pointed out in Henrich & Boyd 2001). They show that there is an important asymmetry between altruistic cooperation and altruistic punishment because, in the absence of altruistic punishment, within-group selection against altruistic cooperation is always strong. Thus, cultural group selection cannot sustain altruistic cooperation, without altruistic punishment.

Despite their formal rigor, cultural group selection models are often treated with skepticism because of the long running controversy about genetic group selection within biology (Sober & Wilson 1998). However, this skepticism, where it is found, is typically based on over generalized suspicions of anything using the word 'cultural' or 'group selection', and not an in-depth understanding of the details of the mathematical *differences* between cultural evolution (and culture-gene coevoution) and genetic evolution. The criticisms that apply to, and make genetic group selection an unlikely force in human evolution, do NOT apply to cultural and culture-gene co-evolutionary models discussed above (Boyd and Richerson, in press; Henrich & Boyd 2001; Henrich forthcoming). As we found at the recent Dahlem conference, skeptics are unable to come up with a specific criticisms, and when pressed, they voice only an untargeted, vague distrust.

Gintis (2000) also developed an evolutionary model showing how strong reciprocity can evolve and persist in evolutionary equilibrium.[9] His model is based on the plausible idea that in the relevant evolutionary environment human groups faced extinction threats (wars, famines, environmental catastrophes) with a positive probability. When groups face such extinction threats, neither reciprocal altruism nor indirect reciprocity can sustain the necessary cooperation that helps the groups to survive the situation because the shadow of the future is too weak. Kin-selection also does not work here because in most human groups membership is not restricted to relatives but is also open to non-kin members. However, groups with disproportionately many strong reciprocators are better able to survive these threats. Hence, within-group selection creates evolutionary pressures against strong reciprocity because strong reciprocators engage in individually costly behaviors that benefit the whole group. In contrast, between-group selection favors strong reciprocity because groups with disproportionately many strong reciprocators are

[9] For other evolutionary models of strong reciprocity see Bowles and Gintis (2001) and Sethi and Somananthan (2001a, 2001b).

better able to survive. The consequence of these two evolutionary forces is that in equilibrium strong reciprocators and purely selfish humans coexist. This logic applies to genes, cultural traits or both in an interactive process. Thus, this approach provides a logically rigorous argument why we observe heterogeneous responses in laboratory experiments.

SUMMARY

The main purpose of this paper is to address the question whether strong reciprocity results from the maladaptive operation of a psychology that evolved in ancestral human environments under processes described by the canonical models of cooperation (reciprocal altruism, indirect reciprocity and reputation). The weight of the evidence suggests that strong reciprocity is unlikely to be accounted for by the 'maladaptationist approach'. This means that the prevailing evolutionary accounts, which often ignore population structure and our second system of inheritance (culture), cannot explain strong reciprocity. Cultural group selection models and culture-gene co-evolutionary model are capable of providing ultimate equilibrium explanations of strong reciprocity. We do not believe that these models are the last word in the debate about the ultimate causes of strong reciprocity. More empirical and theoretical work is necessary to evaluate the plausibility of these models and to discriminate between them and possibly other forthcoming accounts of strong reciprocity. Because of our limited empirical knowledge about human evolutionary history, and the general lack of systematic attempts by evolutionary theorists to generate sharp testable predictions, our conclusions necessarily have to be preliminary. In the future, we strongly encourage theoreticians to generate sharply focused, testable, hypothesis that allow discrimination between alternative ultimate explanations. Empirical work should not just aim at providing evidence that is consistent with one of the prevailing approaches but should also aim at discriminating between competing approaches. There is thus ample room for further theoretical and empirical investigations.

REFERENCES

Alexander, R.D. 1987. The Biology of Moral Systems. New York: Aldine De Gruyter.

Arnold 1996. The Archaeology of Complex Hunter-Gatherers. *Journal of Archaeological Method and Theory* 3(2): 77-126

Aschmann, H. 1959. The Central Desert of Baja California. Ibero-Americana 42. Berkeley: Univ. of California.

Binford, L. 1983. In Pursuit of the Past. London: Thames and Hudson.

Bowles, S., and H. Gintis. 2001. The evolution of strong reciprocity. Discussion Paper, Univ. of Massachusetts at Amherst.

Boyd, R., H. Gintis, S. Bowles, and P. Richerson. 2002. The evolution of altruistic punishment. *Proc. Natl. Acad. Sci*., in press.

Boyd, R., and P.J. Richerson. 2003. Solving the puzzle of human cooperation. In: Evolution and Culture, ed. S. Levinson. Cambridge MA: MIT Press, in press.

Camerer, C.F., and R.H. Thaler. 1995. Ultimatums, dictators and manners. *J. Econ. Persp.* **9**:209–219.

Cameron, L.A. 1999. Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Econ. Inq.* **37(1)**:47–59.

Cane, S. 1990. Desert demograhy: A case study of pre-contact aboriginal densities. In: Hunter-Gatherers Demography: Past and Present, ed. B. Meehan and N. White, pp. 149–59. Oceania Monograph 19. Sydney: Univ. of Sydney.

Case, A., I. Lin, and S. McLanahan. 2001. Educational attainment of siblings in stepfamilies. *Evol. Hum. Behav.* **22**:269–289.

Daly, M., and M. Wilson. 1988. Homicide. New York: Aldine De Gruyter.

Falk, A., E. Fehr, and U. Fischbacher. 2001. Driving forces of informal sanctions. Working Paper No. 59. Institute for Empirical Research in Economics, Univ. of Zurich.

Falk, A., and U. Fischbacher. 1999. A theory of reciprocity. Working Paper No. 6. Institute for Empirical Research in Economics, Univ. of Zurich.

Fehr, E., and U. Fischbacher. 2002a. Retaliation and reputation. Institute for Empirical Research in Economics, Univ. of Zürich, mimeo.

Fehr, E., and U. Fischbacher. 2002b. Why social preferences matter: The impact of non-selfish motives on competition, cooperation and incentives. *Econ. J.* **112**:C1–C33.

Fehr, E., U. Fischbacher, and S. Gächter. 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum. Nat.* **13**:1–25.

Fehr, E., and S. Gächter. 1998a. How effective are trust- and reciprocity-based incentives. In: Economics, Values and Organization, ed. A. Ben-Ner and L. Putterman, pp. 337–363. Cambridge: Cambridge Univ. Press.

Fehr, E., and S. Gächter. 1998b. Reciprocity and economics: The economic implications of homo reciprocans. *Eur. Econ. Rev.* **42**:845–859.

Fehr, E., and S. Gächter. 2000. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**:980–994.

Fehr, E., and S. Gächter. 2002. Altruistic punishment in humans. *Nature* **415**:137–140.

Fehr, E., S. Gächter, and G. Kirchsteiger. 1997. Reciprocity as a contract enforcement device. *Econometrica* **65**:833–860.

Fehr, E., E. Tougareva, and U. Fischbacher. 2002 Do high stakes and competition undermine fairness? Working Paper No. 125. Institute for Empirical Economic Research, Univ. of Zurich.

Fischbacher, U., C. Fong, and E. Fehr. 2002. Competition and fairness. Working paper No. 133. Institute for Empirical Research in Economics, Univ. of Zurich.

Fischbacher, U., S. Gächter, and E. Fehr. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* **71**:297–404.

Gächter S., and A. Falk. 2002. Reputation and reciprocity: Consequences for the labor relation. *Scand. J. Econ.* **104**:1–25.

Gintis, H. 2000. Strong reciprocity and human sociality. *J. Theor. Biol.* **206**:169–179.

Gintis, H., E. Smith, and S. Bowles. 2000. Costly signaling and cooperation. *J. Theor. Biol.* **213**:103–119.

Güth, W., R. Schmittberger, and B. Schwarze. 1982. An experimental analysis of ultimatium bargaining. *J. Econ. Behav. Org.* **3**:367–88.

Hamilton, W.D. 1964. Genetical evolution of social behavior I, II. *J. Theor. Biol.* **7(1)**:1–52.

Harpending, H. 1998. Regional variations in !Kung popuations. In: Kalahari Hunter-Gatherers: Studies of the !Kung San and Their Neighbors, ed. R.B. Lee and I. DeVore, pp. 153–165. Cambridge, MA: Harvard Univ. Press.

Henrich, J., and R. Boyd. 2001. Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *J. Theor. Biol.* **208**:79–89.

Henrich, J., R. Boyd, S. Bowles et al. 2001. In search of homo economicus: Behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* **91**:73–78.

Hoffman, E., K. McCabe, and V. Smith. 1996. On expectations and monetary stakes in ultimatum games. *Intl. J. Game Theory* **25**:289–301.

Johnson and Earle 2000. The Evolution of Human Societies: From Foraging Group to Agrarian State. Stanford CA: Stanford University Press.

Keely, L.H. 1996. War before Civilization. New York: Oxford Univ. Press.

Kelly, R. 1995. The Foraging Spectrum. Washington, D.C.: Smithsonian Institution Press.

Keser, C. and F. van Winden. 2000. Conditional Cooperation and Voluntary Contributions to Public Goods. Scand. J. of Econ. **102**:23-39.

Lee, R.B. 1998. !Kung spatial organization: An ecological and historical perspective. In: Kalahari Hunter-Gatherers: Studies of the !Kung San and Their Neighbors, ed. R.B. Lee and I. DeVore, pp. 74-97. Cambridge, MA: Harvard Univ. Press.

Manson, J.H., and R.W. Wrangham. 1991. Intergroup aggression in chimpanzees and humans. *Curr. Anthro.* **32**:369–390.

McCabe, K.A., S.J. Rassenti, and V.L. Smith. 1998. Reciprocity, trust, and payoff privacy in extensive form bargaining. *Games Econ. Behav.* **24**:10–24.

Milinski, M., D. Semmann, and H.-J. Krambeck. 2002. Reputation helps solve the "tragedy of the commons." *Nature* **415**:424–426.

Myers, F.R. 1986. Pintupi Country, Pintupi Self: Sentiment, Place and Politics among Western Desert Aborigines. Washington, D.C.: Smithsonian Institution Press.

Nowak, M.A., and K. Sigmund. 1998. Evolution of indirect reciprocity by image scoring. *Nature* **393**:573–577.

Peterson, N. 1975. Hunter-gatherer territoriality: The perspective from Australia. *Am. Anthro*. **77**:53–68.

Peterson, N. 1978. The importance of women in determining the composition of residential groups in aboriginal Australia. In: Woman's Role in Aboriginal Society, ed. F. Gale, pp. 16–27. Canberra: Australian Institute of Aboriginal Studies.

Peterson, N., and J. Long. 1986. Australian Territorial Organization. Oceania Monograph 30. Sydney: Univ. of Sydney.

Price, M.E., L. Cosmides, and J. Tooby. 2002. Punitive sentiment as an anti-free rider psychological device. *Evol. Hum. Behav.* **23**:203–231.

Rabin, M. 1993. Incorporating fairness into game theory and economics. *Am. Econ. Rev.* **83**:1281–1302.

Roth, A.E. 1995. Bargaining experiments. In: Handbook of Experimental Economics, ed. J. Kagel and A. Roth, pp. 253–348. Princeton: Princeton Univ. Press.

Roth, A.E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir. 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *Am. Econ. Rev.* **81**:1068–1095.

Rozin, P., L. Millman, and C. Nemeroff. 1986. Operation of the laws of sympathetic magic in disgust and other domains. *J. Pers. Soc. Psych.* **50**:703–712.

Sethi, R., and E. Somananthan. 2001. Preference evolution and reciprocity. *J. Econ. Theory* **97**:273–297.

Sethi, R., and E. Somananthan. 2003. Understanding reciprocity. *J. Econ. Behav. Org*. **50(1)**:1–27.

Silk, J.B. 1980. Adoption and kinship in Oceania. *Am. Anthro.* **82***:*799–820.

Silk, J.B. 1987. Adoption and fosterage in human societies: Adaptations or enigmas? *Cult. Anthro*. **2**:39–49.

Silk, J.B. 2002. The evolution of cooperation in primate groups. In: The Foundations of Social Reciprocity, ed. H. Gintis, R. Boyd, S. Bowles, and E. Fehr. Princeton: Princeton University Press, in press.

Slonim, R., and A.E. Roth. 1998. Financial incentives and learning in ultimatum and market games: An experiment in the Slovak Republic. *Econometrica* **65**:569–596.

Sober, E. and D. S. Wilson 1998. Unto others – the evolution and psychology of unselfish behavior. Cambridge, Mass.: Harvard University Press.

Thomson, Donald F. 1932. Ceremonial Presentation of Fire in North Queensland: A Preliminary Note on the Place of Fire in Primitive Ritual. *Man* 32: 162-166.

Tomasello, M., and J. Call. 1997. Primate Cognition. Oxford: Oxford Univ. Press.

Trivers, R. 1971. The evolution of reciprocal altruism. *Qtly. J. Biol.* **46**:35–57.

Weissner, P. 1982. Risk, Reciprocity and social influences on !Kung San economics. In:Politics and History in Band Societies, ed. E. Leacock and R. Lee. New York: Cambridge University Press.

Zahavi, A., and A. Zahavi. 1997. The Handicap Principle: A Missing Piece of Darwin's Puzzle. New York: Oxford Univ. Press.