Chapter from:
Genetic and Cultural Evolution of Cooperation
Edited by P. Hammerstein, 2003
ISBN 0-262-08326-4 © The MIT Press

7

Group Report: The Role of Cognition and Emotion in Cooperation

Richard McElreath, Rapporteur
Timothy H. Clutton-Brock, Ernst Fehr, Daniel M.T. Fessler,
Edward H. Hagen, Peter Hammerstein, Michael Kosfeld,
Manfred Milinski, Joan B. Silk, John Tooby, and Margo I. Wilson

INTRODUCTION

Altruism, behavior which reduces the individual fitness of the actor while increasing the fitness of another organism, has attracted much attention from both biologists and economists because it seems to defy the logic of both natural selection and standard preferences. In biology, kin selection (Hamilton 1964) is the best-established explanation of the evolution and maintenance of altruistic behavior. However, many examples of apparent altruism defy explanation by kin selection, since they occur among unrelated individuals. The second best-established theory, reciprocal altruism (Trivers 1971), offers to explain substantial portions of this remainder. However, outside of humans, little good evidence exists, so its status is still undetermined. In addition, many examples of putative altruism in humans, particularly those of greatest interest to economists, defy explanation by reciprocal altruism, either because they occur within very large groups of individuals or occur without the possibility of reciprocation. Thus the challenge before us is to understand better the range of mechanisms that support cooperation, particularly outside kin selection.

In this chapter, we summarize our discussions of mechanisms that support altruism outside of kin selection. We felt it was important to focus our discussion on mechanisms. One of the strengths of Darwin's account of adaptations is that it not only explains why animals are often well-adapted to their environments, but also why they are often poorly adapted. If all Darwinism did was to predict that animals should be well-adapted, its predictions would be indistinguishable from Creationism. Instead, the theory of natural selection provides a mechanism by which adaptations as well as maladaptations are constructed. It is in this way that attention to mechanisms in the study of cooperation is scientifically productive. A model of cooperation that focuses only on outcome cannot easily predict when cooperation does not emerge. Simultaneously, without attention to errors in the functioning of cognitive machinery or flaws in specific algorithms, we may not be able to understand the design of the machinery we do find. Although the distinction between mechanism (proximate explanation) and function (ultimate explanation) is useful, it obscures the modern understanding that mechanisms have strong impacts on function.

Economists, like biologists, have been interested in the emergence and stability of cooperative behavior. They also have good reason to turn to mechanisms as assets in designing both models and experiments. A substantial body of experimental evidence now confirms that human behavior substantially deviates from the predictions made by standard models of selfish rationality. However, this confirms only that people do not have standard preferences, that their utilities do not emerge in a simple way from the explicit payoffs. Behavioral economics has emerged as a way of uniting traditional tools with a concern for dissecting the components of the utility functions behind economic theory, as well as exploring alternatives to optimizing strategies. These debates must focus on the details of how individuals, for example, infer intention and compute concepts such as “fairness.” The specific form which rationality takes, the nature of algorithms in an individual’s head, and the cues which individuals attend to and how they use them, all influence behavior in potentially cooperative settings.

This report is organized as follows. First, we discuss evidence for reciprocal altruism in animal societies, as well as specific mechanisms for the bookkeeping of past interactions. Next, we explore the role of reputation and strong reciprocity in dyadic cooperation. After these two sections on dyads, we discuss the role of reciprocal altruism, strong reciprocity, and reputation for cooperation in sizable groups of individuals, not just pairs. Finally, emotions may as well implement strategies in both dyadic and large-scale cooperation, and the nature of emotion mechanisms may powerfully affect our behavioral predictions in any of these contexts.

BOOKKEEPING

“Do unto others as they do unto you” is not quite the Golden Rule, but it is in the theory of reciprocal altruism. Trivers (1971) brought biologists’ attention to the possibility of altruism contingent upon the altruism of other individuals. Axelrod’s (1984) tournaments and Axelrod and Hamilton’s (1981) model of reciprocal altruism went a long way toward popularizing the prediction that cooperation in pairs of unrelated individuals could be sustained if individuals (a) recognize one another, (b) individuals keep track of past interactions, and (c) contingently help those who helped in the past. Consequently, the "keeping track," or bookkeeping, of past interactions has been the focus of much work on reciprocal altruism, much as kin recognition has been in kin selection. We begin by reviewing the empirical evidence that bookkeeping allows unrelated animals to sustain cooperation. We then present theory and observations about the nature of bookkeeping strategies in dyads which suggest that, in some contexts, careful bookkeeping may not always be such a clear prediction after all.

Evidence of Bookkeeping in Nature

Outside of humans, good evidence of reciprocal altruism is quite limited. Hammerstein (Chapter 5, this volume) discusses significant flaws with several of the most widely cited studies of reciprocal altruism in nonhuman animals (see also Enquist and Leimar 1993). A number of studies do not explicitly examine contingency of aid. Instead, many studies, including those on nonhuman primates, simply provide correlations between help given and received for particular pairs of individuals (Silk, this volume). The main problem that arises in correlational studies of reciprocal altruism (as in all correlational studies) is that it is difficult to be certain that the association between two forms of behavior is not the product of some third variable that has not been measured. Thus, some researchers have reported a positive correlation between the amount of grooming within dyads and the amount of social support within dyads (Silk, this volume). It is possible that this correlation reflects contingent behavior: "I will continue to groom you as long as you respond to my solicitations for support." However, it is also possible that this correlation reflects a noncontingent preference for certain partners, such as close kin or age mates or familiar associates ("friends"). Correlational data are also problematic because they hide variation across dyads. If noncontingent cooperation among kin is common, then small, but selectively important amounts of reciprocal altruism among nonkin might be difficult to detect in group-level analyses. This would occur if, for example, all but one dyad in a social group were comprised of related individuals who cooperated without need for reciprocal altruism, since kin selection maintains cooperation in these dyads. However, the lone unrelated dyad might be maintained by reciprocal exchanges but vanish in a group-level analysis. Thus aspects of both the positive and negative evidence are still in question.

Experimental studies, in which contingencies are explicitly examined, provide more convincing evidence that individuals keep track of past exchanges and use that information to direct aid selectively, at least in nonhuman primates (reviewed in Silk, this volume). However, even when a study explicitly examines contingency, the evidence can remain unclear. This is because, in naturalistic settings, it is very difficult to detect contingencies in behavior. In vervets and macaques, grooming is linked to subsequent support (or apparent willingness to provide support) in experimental settings (Hemelrijk 1994; Seyfarth and

Cheney 1984); however, grooming is not consistently correlated with support among nonrelatives in naturalistic settings (Schino 2001). Among captive chimpanzees, possessors of food are more likely to share with former groomers than with others and are less likely to behave aggressively to attempts to share by former grooming partners than with others (de Waal 1997). However, the absolute magnitude of the effect of grooming on subsequent grooming is very small; and in dyads that groom often, the contingency disappears. The relevant time interval for judging contingent behavior is still unclear. Reciprocity may be more delayed in the more stable pairings but still maintain cooperation.

The entire literature is, however, not so ambiguous. Ungulates (e.g., impala; Hart and Hart 1992), some rodents (Stopka and Graciasova 2001), and some monkeys (Barrett and Henzi 2001; Cords 2002) exchange grooming reciprocally, taking turns grooming one another. Thus, A grooms B for a short period; then B grooms A; then A grooms B again, etc. In some cases, changes in the length of each grooming sequence within the bout are matched by the other partner. In baboons, however, time matching does not occur in all bouts; roughly 40% of all grooming bouts involve unilateral interactions (A groomed B, but was not groomed by B).

Henzi and Barrett (2002) have presented evidence which suggests that female baboons "trade" grooming for access to other females' newborn infants. In nearly all primate species (including humans), infants are extremely attractive to females other than their mothers. In macaques and baboons, females are quite eager to inspect, greet, and touch other females' infants, but do not hold, carry, or nurse them. Many researchers have noticed that females often use grooming to gain access to infants, but Henzi and Barrett were the first to show that the "price" (grooming time) females pay for access to infants depends on the relative rank of the mother and the handler. Mothers are groomed longer by lower-ranking than by higher-ranking females who want to handle their infants.

Additional evidence from shoaling fish suggests the importance of reciprocal altruism in maintaining cooperative dyads, through both evidence of immediate bookkeeping and the nature of cooperating groups. In the wild, when groups of sticklebacks (*Gasterosteus aculeatus*) have detected a predator, such as a pike, they do not normally flee or hide. Instead, single fish or small groups leave the school and approach the predator very closely, waiting a few moments within striking distance of the predator. One fish moves forward a bit, and if the other one follows, the first proceeds a bit more, perhaps monitoring the partner's continued cooperation. It has been shown experimentally that this behavior is contingent (Milinski 1987). The fish inspect repeatedly with the same partner in a way consistent with a contingent reciprocal strategy. (For a discussion of the controversy surrounding this evidence, cf. Dugatkin 1997.) Usually pairs, but not larger groups, of sticklebacks participate in these so-called predator inspection visits. This may seem puzzling, since the cost of predator inspection would be smaller in larger groups, due to risk dilution. However, theoretical work by

Boyd and Richerson (1988) suggests that reciprocal altruism is unlikely to evolve in large and even moderately sized groups. (This result is explained in a later section.) Among the sticklebacks, even the rarer, larger inspection groups have been shown to consist of several well-synchronized pairs (Milinski et al. 1990), not a large well-synchronized whole. These experiments and observations thus constitute indirect evidence of direct reciprocity, since altruism driven by reciprocity should be confined to small groups of individuals.

Similar evidence from social carnivores makes the same suggestion. Coalitions consisting of two to nine male lions take over groups of females and defend them against male rivals who persistently attempt to overthrow them. These coalitions can hold a group for two years on average, and during this time they father offspring. Defending the group against other lions is a risky altruistic behavior, since males who may defend less benefit from others' defense. Boyd and Richerson's (1988) prediction is fulfilled here as well. Packer et al. (1991) found that while successful coalitions of two or three male lions often consisted of unrelated individuals, larger groups consisted of close kin. One interpretation of these results is that, in the small coalitions, reciprocal altruism could successfully maintain cooperation. In larger coalitions, kinship was instead the only viable option. (Packer has another interpretation of these observations, invoking sharing paternity within the pride.)

In the preceding examples, the actual costs and benefits of the behaviors in question are very unclear. Part of the debate about bookkeeping in nature is about whether each example is indeed an example of altruism. It is very difficult to measure, or even estimate, the costs and benefits of alternative behaviors. Milinski et al.'s (1997) elegant and painstaking experiments with sticklebacks illustrate this point. Only after two years of investment in experimental design were they able to measure the risks associated with inspection behavior precisely. Fish who lag behind (and therefore "defect") are indeed less likely to be taken by the predator, although with a significantly nonzero probability. The probabilities of capture provide estimates of cost parameters and suggest that inspection really is costly to individuals, that closer inspection entails greater costs, and that "defection" reduces these costs. Furthermore, fish do not seem to be engaging in costly signaling of their own quality, as fish which advance further than their partners and then return to the same position are no better at escaping attacks, which casts doubt on one important alternative explanation. Another two years were needed to estimate the benefits of inspection behavior, which seem to be some function of the advantage of feeding in safety when the fish has information suggesting that the predator is not hungry and will not strike.

After all this careful experimental work, we still do not know how well these costs and benefits generalize to the wild, and perhaps because of this, predator inspection remains controversial (Dugatkin 1997). Milinski et al.'s studies illustrate that the lack of convincing evidence for reciprocal altruism in nature is partly due to the difficulty of measuring the relevant costs and benefits, as well as performing the correct contingency tests. Thus we should not yet conclude that the absence of evidence suggests the absence of contingent reciprocal strategies which maintain cooperation in pairs. Further, we think that this situation provides an appealing opportunity for thoughtful and careful empirical studies to make a big impact, whatever the results.

Cooperation without Bookkeeping

There is a conspicuous discontinuity between humans and other animals in the prevalence of reciprocal altruism. It requires no special methodology to demonstrate that human life relies on a series of exchanges among nonrelatives. Every time we pay for our groceries or revise our colleagues' manuscript, we are practicing some kind of reciprocal strategy. However, it is not entirely clear whether the same contingency mechanisms shape all kinds of cooperative dyadic relationships in human societies. Silk (this volume) reviews evidence that friendship in humans violates the contingency and bookkeeping predictions of reciprocal altruism theory. Reviewing a number of studies from social psychology, she argues that the evidence on human friendship suggests that friends do not keep careful accounts. In fact, the apparent or actual absence of bookkeeping is often taken as one of the best signals of friendship. Most of the evidence comes from Western subjects, and so these results may not generalize to most human societies. If they do, evolutionary theorists face the challenge of explaining either how some of the most significant cooperative relationships in humans might function without detailed bookkeeping or why individuals present the image that they are not keeping track.

Most people recall some proportion of interactions in friendships and other reciprocal relationships. We all have intuitions that people recall instances of aid or defection from the distant past, perhaps reciting such lists in angry moments. However, experimental evidence exists which suggests that people may be forgetting or not even bothering to store much more. Milinski and Wedekind (1998) performed an experiment designed to investigate the use of two different bookkeeping strategies in an iterated Prisoner's Dilemma (PD) setting. The first, Pavlov (Nowak and Sigmund 1993), attends to both its own and its partner's previous round payoffs, in deciding how to behave in the present. The second, Generous Tit-for-Tat (GTFT; Nowak and Sigmund 1992), simply copies what its partner did in the last round but sometimes cooperates when its partner defected. Since these two strategies differ in the amount of memory they require (Pavlov needs more), Milinski and Wedekind introduced a memory constraint into the game by requiring subjects to play a game of memory, in which they had to match symbols on the backs of a field of cards. After each round of the PD with a fixed partner, each subject was allowed to turn over two cards. If they did not match, the cards were turned back over. Subjects were told they would be paid the *product* of their scores in the iterated PD and the memory game, meaning a subject could not afford to ignore either game.

The results showed that subjects' behavior was more consistent with a GTFT strategy when under memory constraints, but with a Pavlovian strategy in the absence of those constraints. These results suggest that memory space is really a finite resource and that strategies which keep simple tidy books can therefore outperform those with detailed books, under the right conditions. This calls into question whether it is always practical for people to keep detailed accounts of interactions in long-term cooperative relationships. Instead, they may be tracking only recent interactions, or only interactions with substantial costs and benefits. Currently, we know of no evidence sufficient to answer these questions, since high-quality data on the life histories of human friendships are sorely lacking.

Theoretical work also suggests that strategies which keep more detailed accounts may not be more adaptive, in some environments. Bendor et al. (1991) conducted a computer tournament using a continuous variant of the repeated PD which casts some doubt on the intuition that Tit-for-Tat, like bookkeeping, is a good strategy in all reciprocal interactions. Bendor solicited computer strategies much like Axelrod (Axelrod and Hamilton 1981; Axelrod 1984) did during his tournaments. Strategies were paired at random and played a repeated game. During each round of the game, each player picks a number between zero and one. Larger numbers cost the player more and benefited its partner more. Individuals observed the other player's number, but with normal random error added. Strategies which kept running accounts, and attempted to return as much on average as they received, did badly. Tit-for-Tat also did badly. The strategies that did best were ones that chose a number that was some modest percentage larger than the number they observed their opponent use during the previous period. Bendor argues that account-keeping rules did badly because errors in perception caused them to walk randomly through the space between zero and one. Such strategies over-fit their observations, taking every deviation far too seriously. In contrast, strategies that were a little nicer than their opponent tended to bump up toward the maximum payoff without too much risk of exploitation and were robust in the face of perception errors. Of course, the nature of successful strategies does depend upon the mix of strategies in the population, and thus these results may not be robust. They do, however, suggest that we should be careful about the intuition that only account-keeping strategies can be successful and avoid exploitation.

To understand more fully the mechanisms that sustain dyadic cooperation in humans, we need both more theoretical work investigating the range of environments in which strategies that keep short and (as above) optimistic accounts do well, and more theoretically grounded empirical work investigating the nature of friendship and the ontogeny of cooperative relationships. The experimental and theoretical results above suggest that the optimal amount of bookkeeping may be low, given memory requirements and perception errors. In addition, which interactions one should regard as important for reciprocal altruism remains an open question. If interactions vary in the magnitude of benefits and costs, then attending only to substantial instances in which perception errors will have smaller effects, may be a better strategy than regarding all interactions as equally informative.

REPUTATION IN DYADIC COOPERATION

Although the issues in the preceding section concern dyads keeping track of past behavior, potential cooperators might also be interested in the past behavior of individuals with whom they have not yet themselves cooperated. Most people have a strong intuition that reputation, some index constructed from past social behavior, is important in human cooperation. Alexander (1987) suggested that *indirect reciprocity*, in which third parties either observe or hear about the behavior of members of their social groups, might support cooperation. About the same time, Sugden (1986) developed a small family of models of such a process. Similar ideas about the power of third-party knowledge have also arisen in noncooperative and nonhuman contexts, such as the formation of linear dominance hierarchies (Chase 1982; Chase et al. 2002; also Tomasello and Call 1997) and in animal conflict (Johnstone 2001).

Indirect reciprocity, if it works, must rely upon some distributed bookkeeping system, in which information about past behavior travels through social networks and regulates ongoing cooperative behavior. Boyd and Richerson (1989) modeled one version of Alexander's idea of indirect reciprocity, involving a circular chain of benefits. However, this mechanism supported cooperation under only small and very long-lived associations, much like reciprocal altruism. Although Sugden (1986) worked on the problem earlier and developed a plausible mechanism, it was not until Nowak and Sigmund's (1998a, b) models of indirect reciprocity that much interest in reputation mechanisms reemerged.

In this section, we review the theoretical work on reputation in dyadic cooperation as well as the experimental evidence. It is important to note that reputation in these models does not solve problems of cooperation in large groups. All of the cooperation here happens within dyads. We discuss reputation and other mechanisms which may maintain cooperation in larger groups in a later section.

Image Scoring and Standing

There are two components to any indirectly reciprocal strategy: (a) how the accounts are kept and (b) how the accounts are used to make decisions. Nowak and Sigmund (1998a, b) modeled indirect reciprocity with a system of bookkeeping they call *image scoring*. Image scoring works in the following way. Each individual in a social group is characterized by an image score, which is a positive or negative integer. Whenever an individual has the opportunity to aid another individual, this image score increases by one if he donates aid (cooperates) and decreases by one if he does not donate aid (defects). It is assumed that image scores are completely accurate and common knowledge: every individual knows (or has access to) the image score of every other individual, as well as his own, without error. Nowak and Sigmund then proposed a strategy which discriminates based upon image scores. If a discriminating cooperator is paired with an individual with an image score above a given threshold, the discriminator provides aid (cooperates). Otherwise, the discriminator refuses aid (defects). It is important to note that this strategy is insensitive to the effects of its behavior on its *own* image score. A discriminator of this kind will defect with an individual of low image score, even though that defection reduces her own image score by one unit. In this regard, the image scoring and discriminating strategy is providing altruistic punishment.

Some work demonstrates that image scoring can sustain cooperation. Nowak and Sigmund (1998b) modeled a world of 100 individuals in a single social group. Each generation, individuals were paired at random with one other individual to whom they had the option of providing aid, which was an altruistic act. After behavior, image scores were updated, and each individual was matched with another random individual. There were no fixed cooperating dyads. Nowak and Sigmund found that the discriminator strategy, although it never went to fixation against a pure defection strategy, sustained about a 40% frequency in the group over the long run.

Later simulation work challenges these results, however. Leimar and Hammerstein (2001) became interested in how well the image scoring results would generalize in a more realistic model. Theory always contains an antagonism between realism and tractability. We want theories which capture only the important details, but no more, lest the model become just as incomprehensible as reality. However, Nowak and Sigmund's model contained an assumption that does not fit the problem under study. In their simulations there existed only one social group, of only 100 individuals. Such a population structure is known to result in large amounts of drift, overwhelming selective forces. Furthermore, if we are thinking of a genetic model of human populations, even in the distant past, effective population sizes (N_e) were probably on the order of tens or hundreds of thousands (the low-bound estimate is around 10,000 over the last 1–2 million years; Relethford 1998). There has been some debate about these estimates, but the debates have focused on the probability that current simulations *under*estimate N_e, not that they *over*estimate it (Hey 1997; Wolfpoff 1998).

To see if this assumption of a small lone group made a difference, Leimar and Hammerstein simulated Nowak and Sigmund's image scoring model with a population of 100 groups of 100 individuals each (a maximum N_e of 10,000). Groups were linked by migration, such that when migration was reduced to zero, they could reproduce the Nowak and Sigmund results; with increasing amounts of migration, however, the results differed substantially. With even modest amounts of mixing among groups, image scoring and discrimination began to perform quite badly. The reason is that a complex interaction of powerful drift and selection were driving the cycles of evolution of the image scoring strategy, but in the larger effective population, drift was much weaker and these interactions did not arise.

In a genetic model, image scoring has some serious problems. It should not be overlooked that a model assuming cultural rather than genetic transmission is much less constrained in its assumptions about effective population size. For cultural transmission, Nowak and Sigmund's model might be a reasonable approximation of the dynamics.

A more serious problem with the image scoring strategy, which both genetic and cultural models face, is that it is easily invaded by strategies which Nowak and Sigmund did not consider. Leimar and Hammerstein introduced a strategy which attends only to its *own* image score, ignoring the image score of its partner. If such an individual's image score is above the discriminator strategy's threshold for providing aid, it defects. If its image score is below the threshold or equal to it, it cooperates. Introduced into Nowak and Sigmund's model, this strategy quickly replaces the image scoring and discriminating strategy. The reason is that discriminators help such image score seekers, and the image score seekers take advantage of discriminators.

To solve this problem of invadibility, Leimar and Hammerstein introduced a strategy invented by Sugden (1986) which instead keeps track of *standing*. An individual's standing can be either *good* or *bad*. An individual gains or retains good standing by providing aid to another individual. An individual loses good standing and attains bad standing by failing to aid another individual in good standing. Failing to aid an individual in bad standing, however, does not result in a loss of good standing. These are justified defections. They then considered a strategy, called the standing strategy, which provides aid to individuals with good standing but refuses to aid individuals in bad standing. They found that the standing strategy outperformed the image scoring strategy, even in the presence of execution and perception errors. Nowak and Sigmund (1998a) suggested that standing strategies would be more vulnerable to errors in perception than image scoring strategies. According to Leimar and Hammerstein's simulations, this is probably not true: although errors hurt the standing strategy, it still out-competed image scoring.

Image scoring suffers from two serious deficits: (a) it is exploitable by image-seeking strategies which defect after achieving high image scores and (b) it provides a form of altruistic punishment every time it defects on an individual with a low image score. The results above were produced in the absence of errors in knowledge of reputations. If reputations (i.e., image scores and standings) are known with some error, then image scoring might perform better, since accumulated scores would be less sensitive to random errors than binary standings. However, both strategies must be very sensitive to errors in knowledge (Nowak and Sigmund 1998a), so we await future work to address this question.

Experimental Evidence on Reputation Mechanisms

Theoretical work thus far suggests that standing strategies are more likely candidates for implementations of indirect reciprocity in human societies than are image-scoring strategies. Some of the most recent experimental work disagrees, however. Wedekind and Milinski (2000) showed that groups of eight subjects could sustain cooperation through indirect reciprocity, but these experiments were not designed to distinguish between image scoring and standing strategies. To investigate the specific mechanisms supporting indirect reciprocity, Milinski and colleagues (2001) conducted a series of experiments designed to tease apart image scoring and standing in a simplified indirect reciprocity situation. They set up groups of seven subjects in which one subject was actually a confederate instructed to always refuse to give aid, the "NO" player. Individuals with the opportunity to aid the NO player should refuse to do so whether they are using an image scoring or standing strategy. These strategies should respond differently to refusals to aid the NO player, if given the opportunity to aid players who just had the chance to aid the NO player. Image scorers should refuse to aid the individual who refused to aid the NO player. Individuals using a standing strategy should, however, provide aid to the same individual. The experimenters found that subjects' behavior was better explained by an image scoring than a standing strategy. Furthermore, individuals who refused aid to the NO player seemed to compensate for the damage to their image scores by being more generous to other individuals. Such compensation is hard to explain as a standing strategy, since justified defections would eliminate the need for compensating a defection. This result also hints at a strategy more complicated than the image-scoring strategies explained in the previous section.

Evidence from the Wason selection task (Wason 1968) provides less specific evidence about mechanism, but again suggests that people regulate their behavior toward others contingent upon reputation. (Cosmides [1989] relates the task to reciprocal altruism.) The human brain must serve as the input circuit for reputational memory. To examine the relationship between cheater detection in the Wason task and reputation, John Tooby and colleagues (pers. comm.) conducted experiments in which subjects read descriptions about persons who have the opportunity to cheat, and then either take advantage of the opportunity, or do not. The Wason task measures cheater detection through the proportion of logically correct card selections. If positive reputation information about a person deregulates cheater detection, then we should expect fewer correct card selections in social contract treatments. If negative reputation sharpens cheater detection, we should expect improved performance with the same instrument. The results indicate that prior acts of cheating by a person do not increase cheater detection. However, four refusals to cheat relax cheater detection, but only for that person, suggesting that reputation about specific individuals regulates attention to rule violations on an individual basis.

STRONG RECIPROCITY IN DYADIC COOPERATION

Fehr and Gächter (1998a, b, 2000), Gintis (2000), Henrich and Boyd (2001), Bowles and Gintis (2001), and Fehr et al. (2002) have focused attention on a strategy that differs fundamentally from reciprocal altruism and reputation mechanisms. They have called this strategy *strong reciprocity.* Strong reciprocity applies to two-person interactions as well as to n-person interactions with $n > 2$. A person is a strong reciprocator if she is willing (a) to sacrifice resources to be kind to those who are being kind (= strong positive reciprocity) and (b) to sacrifice resources to punish those who are being unkind (= strong negative reciprocity). The essential feature of strong reciprocity is a willingness to sacrifice resources for rewarding fair behavior and punishing unfair behavior *even if this is costly and provides neither present nor future material 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 (Falk and Fischbacher 1999). Fehr and Gächter (1998a, b) and Fehr et al. (2002) provide experimental evidence indicating that there exist many people who exhibit strong reciprocity and whose existence greatly improves the prospects for cooperation in dyadic as well as in n-person cooperation.

Despite the similarity of terms, it is important to distinguish strong reciprocity from "reciprocal altruism." In one economic (but not necessarily biological[1]) conception of the strategy, a reciprocally altruistic actor is only willing to help another actor if she expects long-term net benefits from the act of helping – call this a forward-thinking reciprocal altruist. In contrast, a strong reciprocator is willing to incur the costs of helping in response to kind acts of the other party even if there are long-term net costs from the act of helping. The distinction between strong reciprocity and forward-thinking reciprocal altruism can most easily be illustrated in the context of a *sequential* Prisoner's Dilemma (PD) that is played *only once*. 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 economic 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 kind. A cooperative act by player A, despite the economic incentive to cheat, is a prime example of such kindness. The kindness of a strong reciprocator is thus *conditional* on the perceived kindness of the other player. In contrast, a forward-thinking reciprocal altruist only cooperates if there are future returns from cooperation. Thus a

[1] Many people have modeled reciprocal altruism with simple rules that lack all consideration of long-term benefits, e.g., Axelrod and Hamilton (1981) and nearly all other evolutionary models of reciprocal altruism. In this case, telling strong reciprocity from reciprocal altruism becomes harder, but is clearly not impossible.

forward-thinking reciprocally altruistic player B will always defect in a sequential *one-shot* PD.

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 colleagues (Fehr and Gächter 1998b; Fehr et al. 1993) describe the results of many generalized 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 40–60% of the B-players typically respond in a strongly reciprocal manner to the choice of player A: Their cooperation reflects player A's cooperation level. If player A chooses zero cooperation, then strongly reciprocal B-players also choose zero cooperation. However, there are also typically between 40–60% of second movers who *always* choose zero cooperation irrespective of what player A does. These players thus exhibit purely selfish behavior.

It is important to emphasize that in all of these experiments, real money (sometimes up to three months' income) was at stake and players remained anonymous before, during, and after the experiment. There was no repeated interaction and the experimental subjects had no chance to build a reputation. Despite the absence of repeated interactions and reputation building opportunities, subjects in the role of player B reciprocated to cooperative actions of player A. Moreover, Gächter and Falk (2002) have shown that if subjects are given the chance to interact repeatedly in the generalized sequential PD, subjects in the role of player B strongly increase their cooperation rate. This was reasonable because in the condition with repeated interactions, player A could punish player B in the next period by ceasing to cooperate with B. The strong increase in the cooperation of player B in the repeated interaction condition suggests that human subjects are well aware of the difference between a one-shot interaction and a repeated interaction and that their choices are conditional on this difference.

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 $2.50, and every dollar invested into punishment of B reduced player B's earnings by $2.50. 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, 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 and punished for low levels of cooperation and, therefore, 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. In the extended version of the sequential PD, in which A can punish or reward, B-players also expect A-players to exhibit strongly reciprocal behavior. This expectation, in turn, causes a large rise in the cooperation of the B-players relative to the situation in which A-players have no opportunity to reward or punish.

MECHANISMS IN *N*-PERSON COOPERATION

Boyd and Richerson (1988) have shown that reciprocal altruism should be confined to small groups of individuals. The theory is complicated in the details, but the intuition behind it is simple. Reciprocal altruists only do well when they are paired with other reciprocators. In all other cases, nasty strategies do better. This is because the only evolutionarily stable strategy in such a game is the one which cooperates only if everyone else cooperates as well. Otherwise, a few defectors will free ride on the efforts of the reciprocators and out-reproduce them. Furthermore, when groups of individuals are large, the chance of getting a group of all reciprocal altruists is very small. Consider, for example, a case in which individuals are grouped together in fives. Even if half of the population consists of reciprocal altruists, the chance of getting five reciprocators in a randomly formed group is 0.5^5, or 0.03. If groups are around twenty individuals or reciprocators are rare, the situation is truly hopeless. The standard solution to this problem is a small amount of assortative group formation, such as kinship. However, assortment will not help in the case of large groups, since the probability of getting a group consisting only of reciprocal altruists falls geometrically with group size. Even if groups are comprised entirely of full siblings ($r = 0.5$), and assuming again that half of the population is cooperators, a group of ten cooperators has a less than 5% chance of forming.[2]

Thus, cooperation that is contingent on cooperation of all other group members is unlikely to be an effective mechanism for cooperation in large groups. This poses a puzzle since humans often cooperate in large groups of unrelated individuals, groups in which benefits cannot be directed to specific individuals but must be disbursed to the entire group. Furthermore, indirect bookkeeping mechanisms discussed earlier do not apply here: indirect reciprocity as described involves pair-wise cooperation, not cooperation in sizeable groups.

In this section, we discuss mechanisms which may support cooperation in larger groups of unrelated individuals, which is sometimes called *n*-person cooperation. We discuss strong reciprocity as well as the role of reputation in the *n*-person setting.

[2] Let p be the frequency of cooperators in the population as a whole. Let groups be comprised of n individuals with an average coefficient of relatedness r. Then the probability of sampling a group of all cooperators is $p \times \{r + (1 - r)p\}^{n-1}$.

Strong Reciprocity in *n*-Person Groups

Cooperation in *n*-person groups is best viewed as a problem of public goods provision. The crucial feature of a public good is that it is difficult or impossible to exclude other group members from the consumption of the good. Hence, those who do not contribute to the production of the good can also consume the good. In the public goods context, strong positive reciprocity means that individuals increase their own contribution to the good if they expect the other group members also to increase their contributions. Strong reciprocators thus condition their choices on the other group members' choices even in one-shot situations. Strong negative reciprocity means that individuals who cooperate are willing to punish those who defected, if given a chance to do so, even if punishment is costly for the punisher and yields no economic benefits whatsoever.

Strong Positive Reciprocity

Fischbacher et al. (2001) examined to what extent strong positive reciprocity is present in one-shot *n*-person public goods situations. 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% 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 the more they expect others to contribute, indicating a strongly reciprocal cooperation pattern. Only 10% of these subjects are willing to match the average contribution of the other group members, whereas 40% of the strongly reciprocal types contribute less than the average contribution of the other group members. Roughly 30% of the subjects behave in a fully selfish manner, always defecting irrespective of how much they expect others to contribute. The rest of the subjects exhibits either a quite erratic contribution pattern (6%) or a hump-shaped pattern (14%).

In Fehr and Gächter (2000, 2002), subjects repeat the public goods experiment over many periods. In each period the subjects choose simultaneously a contribution level. At the end of the period they are informed about the other group members' individual contributions, and then they proceed to the next period to choose again (simultaneously) the contribution level. This is repeated for six periods in total. In each period new groups are formed such that no subject meets another subject twice. This setting ensures that subjects can learn, over time, how to play the game without allowing for repeated interactions. It turns out that the contributions to the public good strongly decline over time, and toward the final period the vast majority of the subjects contribute little or nothing to the public good. This decline in cooperation can be neatly explained by the dynamics of the interaction between strongly reciprocal types and selfish types, as revealed by the results of Fischbacher et al. (2001): For any given expected average contribution of the other group members in period *t*, the 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 expected average contribution in period *t*, inducing the subjects to reduce their expectations about the other members' contributions in period $t + 1$. Due to the presence of reciprocal types, however, 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. Simulations conducted by Fischbacher et al. indicate that the described process captures the actual behavior of the subjects quite well. It is worth emphasizing that a similar decline in cooperation rates is observed in finitely repeated public goods experiments when the group composition remains stable over time. Thus, even if one allows (finitely) repeated interactions between the same people, cooperation cannot be sustained. Despite this decline, cooperation under stable group composition is, in general, higher than when groups are randomly rebuilt every period (see Fehr et al. 2002). This again indicates that subjects understand the difference between one-shot and repeated interactions and behave accordingly.

Note that the Boyd and Richersen (1988) account — why reciprocal altruism cannot explain cooperation in large groups — and the above account — why cooperation in one-shot public goods games cannot be sustained — rely on similar intuitions. Reciprocal altruism cannot flourish in large groups because even a small number of defectors induce a breakdown of cooperation. Likewise, strong positive reciprocity cannot sustain cooperation in one-shot public goods situations because the expectation of even a small number of selfish actors will induce the strongly reciprocal actors to cease to cooperate.

Strong Negative Reciprocity

The previously described public goods experiment is characterized by the absence of targeted punishment opportunities. In this situation subjects can only punish other group members for noncooperation by withdrawing their own cooperation. The withholding of cooperation always punishes all other group members irrespective of whether they contributed or defected. This is the deeper reason for why cooperation cannot be sustained in this setting. The situation changes, however, dramatically if targeted punishment opportunities are made available. This has been done by Fehr and Gächter (2000, 2002) by adding an additional stage at the end of every period. After subjects had made their simultaneous contribution decisions, and after they had been informed about the other group members' individual contributions, each subject in the group had the option of punishing each of the other subjects in the group. Each dollar invested in the punishment of one other group member reduced the income of the punished member by three dollars. When all subjects had made their punishment decisions, they moved to the next period in which they again first chose their contribution levels. The groups were again randomly rebuilt every period so that nobody met anybody else twice.

Selfish subjects will never punish in this situation because punishment is costly and in the future periods they meet only new members. This means that if there are only selfish subjects, the option to target the punishment to specific other individuals in the group is worthless. Since nobody punishes, and since in the absence of targeted punishment nobody has an incentive to cooperate, a group consisting of only selfish subjects will exhibit no cooperation. Strong reciprocators will, however, be willing to punish despite the costs because they view little or no cooperation as an unkind act that deserves to be punished. In fact, a majority of the subjects punished the defectors, and those who were punished increased their contributions in the next period. The existence of targeted punishment led to dramatic changes in overall contribution behavior. Already in the first period of the treatment with targeted punishment, cooperation rates were much higher than in the absence of targeted punishment. Moreover, whereas cooperation unraveled in the absence of targeted punishment, cooperation increased over time when targeted punishment was possible. This indicates that strong negative reciprocity can be a powerful mechanism for obtaining and maintaining cooperation in *n*-person groups.

Fehr and Gächter (2000) also conducted experiments with targeted punishment when the group composition remained stable over (finitely) many periods. Under these conditions it was possible to reach almost 100% cooperation, although in the presence of only self-interested actors the prediction is zero cooperation. Note that in the presence of a stable group composition, the punishment of other group members constitutes a second-order public good because the punished member will in general increase cooperation in the next period and all group members benefit from this increase. It is, therefore, important to distinguish this kind of punishment from punishment in which there is no public goods dilemma. This is the case in two-party interactions (see Clutton-Brock and Parker 1995), where the second-order dilemma is absent.

In view of the powerful effects of strong reciprocity on human cooperation, it is important to develop evolutionary models explaining this phenomenon. Gintis (2000) and Henrich and Boyd (2001) have developed models showing that strong reciprocators persist in evolutionary equilibrium. The challenge for these models is that in the presence of a mix of selfish and cooperative (but nonpunishing) players, those who cooperate and do not punish will do better than those who cooperate and punish because the latter bears the costs of punishing the defectors. However, these evolutionary scenarios remain controversial because they rely on group selection arguments. Chapters 19–23 (this volume) explore in more detail the theory of the evolution of punishment in large groups. An important question for future work is to examine the empirical plausibility of these group selection accounts. Another important yet unsolved question is whether the heterogeneity of behaviors observed in laboratory experiments concerns stable personality differences. Is there such a thing as a strong reciprocator and a selfish type, or do the same subjects sometimes exhibit strongly reciprocal behavior and sometimes purely selfish behavior? How stable are the propensities to reciprocate across time, different games, and different contexts? We are unaware of any good data which address these questions, providing an opportunity for interesting future work.

Reputation and *n*-Person Cooperation

Milinski et al. (2002) studied whether the insertion of reputation in public goods games through interaction with indirect reciprocity games can maintain *n*-person cooperation. They tested this idea with groups of six subjects each. By alternating rounds of a public goods game and an indirect reciprocity game, they found that contributions in the public goods game were maintained at a high level. The results suggest that the need to maintain reputation for the indirect reciprocity game maintained contributions to the public good. However, if subjects no longer expected rounds of indirect reciprocation, contributions to the public good quickly dropped to typically low levels. Thus reputation can maintain cooperation in a public goods game at a level similar as in the punishment experiments of Fehr and Gächter (2000, 2002). Reputation has been shown to raise cooperation levels in subsequent direct reciprocity games also, probably because it builds up trust (Wedekind and Braithwaite 2002).

EMOTIONS

One view of emotions popular in the social and biological sciences is that emotions should be invoked to explain deviations from the norms of rationality. Loewenstein's (1996) work on hot and cold cognition, for example, provides compelling evidence that emotional states affect cognition, although the discussion and experimental design are framed in ways that emphasize the maladaptive consequences of their effects. One gets the impression from much work in these traditions that we would all be better off without emotions. Another view, held in different forms by psychologists in the tradition of Herbert Simon's bounded rationality, evolutionary psychologists, and many others, is that emotions are inseparable and adaptive parts of human decision-making, not forces which necessarily lead us astray. These views suggest ways in which emotion mechanisms process information, together with the more traditionally "cognitive" parts of cognition, to produce adaptive decisions in the real world or environments relevant to the design of human cognition.

We use "emotions" here to refer to a wide category of things people commonly call "feelings." Emotions may prune decision trees, direct attention to specific aspects of the environment, and even prevent our more conscious cognitive apparatus from causing us harm. For example, territorial spiders locked in combat are much easier to approach than those not locked in combat. Attention is a finite resource for any organism, and it is easy to see how focusing on one's opponent, in a situation in which one can die in a few seconds, is an adaptation, not purely a cognitive constraint. Fear in humans probably serves a similar function by directing attention to specific threats. Similarly, Bechara, Damasio, and colleagues (Bechara et al. 1994, 1997, 2000; Damasio 1994) have shown how emotions may be eminently cognitive, weighing probabilities in so-called "multi-arm bandit" tasks. They had normal and brain-damaged subjects participate in a card-stack task. In such tasks, the subject has between two and four stacks of cards, face down, in front of him. He may turn over the card on the top of any stack. In doing so, he receives the payoff printed on the face of the card. Card stacks vary in their expected payoffs, as well as their variances. This task continues for many rounds. During this time, individuals slowly converge on the stack with the highest expected payoff, although this choice behavior seems driven more by impression of "good" and "bad" stacks than conscious understanding of payoff differences. However, some brain-damaged subjects who exhibit low affect never converge on the highest payoff stack nor do they display anticipatory skin reactions of risky choices (as do normal subjects). Even in cases in which brain-damaged subjects developed accurate feelings of "good" and "bad" stacks, they failed to make choices accordingly. These results suggest that emotions play an important information processing role.

Another key feature of emotions is that they are sometimes not penetrable by other parts of cognition. Rozin et al. (1986) 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 knows consciously that the substance is the same fudge they have already eaten, most subjects refuse to eat the fudge in the shape of feces. One interpretation of this and similar experiments (there are many; e.g., Rozin et al. 1986) is that the cues which prime disgust — one of the emotions that regulate consumption — operate independently of other cues. Thus disgust's power over behavior is strong enough such that propositional knowledge that the "dog feces" is really fudge cannot penetrate, leading subjects to forgo a benefit. Although this example might be interpreted as maladaptive behavior on the part of the subjects, it is easy to see how it illustrates adaptive design: in a broad range of environments, objects which resemble feces are not good to eat. Since information about the exceptions is likely difficult to acquire, relying upon a simple set of cues (color, shape) may be more adaptive on average than bothering to learn about each possible food, when the costs of a mistake are likely quite high. Contrived experiments can always make subjects and their cognitive mechanisms look foolish, and we think there is little harm and much more promise in searching for cogent adaptive explanations to be refined and tested.

In this final section, we report on several avenues for exploring emotions as mechanisms that support cooperation in humans. We limit discussion to humans not because of any species prejudice about emotion or its importance in cognition and behavior but rather because data on emotions in nonhuman animals is quite sketchy. We think, however, that the issues explored here suggest ways to investigate the impact of the analogs of human emotions in other animals.

Emotion Mechanisms for Supporting Cooperation

Fessler (1999; Fessler and Haley, this volume) discusses the roles of human emotions in supporting cooperative institutions. One key emotion implicated in cooperative strategies seems to be anger. Cooperative individuals respond with anger to the noncooperative behavior of others, and this appears to motivate them to inflict costs on these defectors. Experiments also find that potential defectors typically anticipate these angry responses (Fehr and Gächter 2002). Thus anger may instantiate part of the mechanisms which lead to strong reciprocity. Also of interest are the eminently normative emotions of shame and pride. Unlike guilt, shame appears to be a human universal and may motivate compliance to norms, including norms which regulate prosocial behavior. Pride is the positive pole of this experience and may function to provide subjective rewards for norm adherence, just as shame provides subjective punishment. Fessler (1999) lays out an evolutionary argument for the function of these emotions in cooperation. Barr (2001) has found that shame can motivate cooperation in experimental games. Bowles and Gintis (this volume) also discuss the role of emotions in regulating cooperative behavior.

Recent evidence using the Wason selection task also suggests that the emotional state is a key part of the instantiation of cooperative strategies. Chang (2002) had subjects complete a mood induction exercise for a specific emotion before completing the social contract version of the Wason selection task (Cosmides 1989). Subjects who successfully completed negative mood induction exercises were significantly better at cheater detection than those who completed positive mood induction exercises (63% vs. 34% correct card selections, respectively). The performance in the negative mood case is similar to usual social contract conditions. However, the positive mood situation led to significantly lower performance than is the norm. This effect of emotional state provides additional evidence that emotions can either direct or deregulate an individual's attention to specific kinds of information or disengage information processing related to cooperative strategies. These behavioral results echo the suggestions of other work by Fehr and Gächter (2002), who found that punishment in a public goods game was motivated by anger, as indicated by subjects' self-reports.

Emotions and Honest Signals

Economists, political scientists, and biologists have long been interested in commitment problems. In many game theoretic situations with sequential play, in which one player moves before the other, the first player has the advantage and gets her way, since the first move restricts the payoffs available to the second player. The second player, however, can grab the strategic advantage if she can "burn her bridges" such that she is constrained to choose an option that is unattractive to the first player. This can be accomplished by really burning one's bridges or by providing credible signals that one is committed to an option. For example, in animal contests, the costs of escalated fights often exceed the value of the resource under dispute. By attacking, a first mover can therefore force a second into retreating from a resource, since it is would be more costly for the second to fight than to flee. If, however, the second animal can commit itself to retaliate any aggression, the first no longer gets a higher average payoff by attacking. Similarly, in situations in which individuals are willing to cooperate if they can be assured that the second player will also cooperate, commitment on the part of the second player can be adaptive.

Signals of intent from the second player are one solution. The trouble, however, is in keeping such signals honest. One puzzling fact about human emotions, unlike the emotions of other animals, is that many are linked to species-typical, fixed, and involuntary facial expressions. Although chimpanzees have some seeming analogs of fixed expressions which correspond to probable emotions, the human repertoire is vast in comparison. Some explanation of this fact is required. It is possible that other animals have similar signals which are olfactory. Whether this is the case or not, some explanation of what exactly these emotions and their expressions are signaling is needed.

Frank (1988), among others, has suggested that involuntary emotional states can help cooperators coordinate by providing solutions to the commitment problem. However, why would natural selection not favor individuals who could fake emotional displays and therefore exploit cooperators? One possibility is that the production of emotional displays is physiologically costly. However, no careful and accepted argument exists as to why this might be the case. Also, for a costly signaling argument, what is important is that the signal be *more* costly for the liar than the honest signaler. Cost alone will not suffice to evolve an honest signal. A careful argument along these lines may be possible, but to our knowledge has not yet emerged in the literature on emotion.

One requirement that all such theories must face is: if there is supposedly a simple and easy-to-evolve signaling mechanism supporting cooperation, then we are left with the mystery of why other animals, and especially other primates who have rich social lives and highly analogous and probably homologous emotions, have not evolved it. One possibility is that smaller-scale primate societies have less opportunity to benefit from cooperation; thus they may have evolved similar mechanisms, but on a smaller scale. However, other primates (e.g., hamadryas baboons) sometimes live in quite large social groups, as large or larger than many human foraging groups. In additional, the size of cooperating groups is partly a result of the evolution of cooperation mechanisms and therefore cannot be easily regarded as an inert exogenous variable.

Given the existence of individuals such as actors and actresses who can convincingly manipulate the overt expression of their emotions, it is worth considering the possibility that natural selection could lead to the ability to fake emotions but that there is some other reason that such lying would not be advantageous in the long run. A problem with our intuitions about signaling equilibria is that almost all models of signaling in animals involve one-shot games. Many people are convinced that honest signals in situations in which animals have at least partly conflicting interests require costly displays or are otherwise simply revealing or unfakeable due to constraints. Silk et al. (2000) have recently provided a simple and intuitive model which explains how honest cheap signals can evolve among unrelated individuals even when interests conflict. The key is to allow repeated interaction and reputation formation. In species as diverse as sparrows and baboons, interactions with the same individuals are often repeated. Silk et al. were inspired by the existence of apparent low-cost and honest signals of intent in a variety of nonhuman species that live in stable social groups. The appropriate contrast, of course, is not between one-shot and repeat interactions but between low and high probability of continuing interacting. Their model shows that high probabilities of continued interaction may drastically change our intuitions about what sorts of signals we should expect to find in nature.

Maynard Smith (1991, 1994) has shown that honest low-cost signals can evolve when interests of individuals are at least partly aligned; they must order the payoffs in the same way. However, these and similar results arise from models which assume that individuals interact only once. Introducing repeat interaction and a memory for events of deception (a signaling reputation of a sort) changes the conclusions. Honest cheap signals can evolve in repeat interactions where they would not be stable in finite relationships. Human emotion displays may have a similar character. Additionally, Farrell and Rabin (1996) have demonstrated that honest cheap signals can be stable when there are substantial conflicts of interest, even in a one-shot game, provided that parties have sufficient incentive to coordinate with one another. An appreciation of these two results, the effects of repetition and coordination, should lead to new ideas about the nature of emotional signals.

Depression as a Bargaining Strategy

Future models of human sociality need to incorporate strategies beyond reciprocity and signaling. In particular, when a cooperative strategy ceases to provide fitness benefits for one of the participants in a cooperative venture, she may find it advantageous to attempt to renegotiate the terms of the venture. Hagen (this volume) proposes that the symptoms of clinical depression — such as loss of interest in virtually all activities — might be elements of a bargaining strategy: an individual who has suffered a serious social loss withholds the benefits she is providing to other group members until they agree to improve the terms of her "social contract." This theory, based on a review of the empirical evidence on clinical depression in Western and non-Western cultures, explicitly links emotions, signals, and bargaining theory to challenge the prevailing view of unipolar depression as a pathology.

Error Management and the Design of Emotion

In reviewing Bendor's (1991) results about the evolution of reciprocity in a stochastic environment, we saw that errors can affect the adaptive design of mechanisms, at least in principle. At the broadest level, emotions, being the product of natural selection, can be expected to reflect the same principal of error management that is to be biased or weighted in such a fashion that, if errors are to occur, they are more likely to be of the sort that, under ancestral conditions, were less rather than more costly (Buss 2001; "error management," Haselton and Buss 2000; Nesse 2001; "smoke-detector principle," Williams and Nesse 1991). The design of disgust, the emotion which guards against contamination (Rozin et al. 1986), may an be an example of error management, because it appears to be elicited when merely superficial cues suggest that contamination is possible. For example, people refuse to eat fudge shaped like feces. Note that error management is operating primarily in the initial interpretation-of-the-stimulus phase of the emotion process (i.e., "Is this fudge or feces?").

By the same token, it is reasonable to expect that error management may affect subjects' interpretation of the tasks they are asked to perform in experimental situations. The interpretation of the "meaning" of cues from the environment is part and parcel of the experience of an emotion (”Is that a shadow in the woods or a jaguar?"). Because the costs of mistaking an iterated game for a one-shot game may have been greater than the costs of the reciprocal error, it is possible that players in one-shot games (particularly when cues are ambiguous) experience emotions appropriate to iterated games and behave accordingly. Except when the format of an experimental game closely matches a familiar cultural practice (Henrich et al. 2001), subjects may experience the game context as somewhat alien, hence calling for interpretation. This interpretation is likely to be subject to the influence of error management effects that stem from both the evolved predispositions and the repertoire of experience. Thus, it is possible that subjects react with anger to perceived transgressions (e.g., inequitable divisions in one-shot ultimatum games) and with shame to perceived disapproval (as with verbal punishment in commons games; Barr 2001) despite the fact that both anger and shame have utility primarily in long-standing interactions.

There is, however, also a competing interpretation of these emotions which stresses that interactions with low probabilities of future encounters have been quite frequent in evolutionary history (see Fehr and Henrich, this volume, and Gintis 2000). In addition, the costs of mistakenly treating an encounter with a low or zero probability of future interactions as an event with a high probability of future interactions may have been quite dangerous so that individuals who were able to distinguish cognitively and emotionally between low- and high-frequency interactions had better survival chances. For instance, treating a stranger like a friend may have been quite costly because it enabled the stranger to exploit the situation and cheat, whereas being cheated by a friend is constrained by the implicit threat of withholding future cooperation. In fact, most modern humans well understand that the probability of being cheated in one-shot interactions in a foreign town or country is higher than in interactions with colleagues and friends. This capacity to distinguish low- from high-frequency encounters, and to behave accordingly, is also documented in the experiments of Gächter and Falk (2002) and Fehr et al. (2002). The competing view is also more optimistic about the human capacity to have emotions that are fine-tuned to low- and high-frequency interactions. Most people probably experience more anger when cheated by a close friend than when cheated by a stranger because the feelings of betrayal tend to be stronger when cheated by a friend.

It is a well-established fact that a substantial fraction of humans cooperate with unrelated strangers even if the shadow of the future or the possibility to build a reputation is absent. Whether the emotions that help sustain cooperation in these low-frequency encounters are ill- or well-adapted to the low-frequency situation is an important topic for future research. We need to know more about subjects' actual default assumptions when they are in one-shot encounters in the laboratory and about the cues that affect the default assumptions. We also need to know more about the details of our evolutionary history, about the likelihood of low-frequency interactions, and about the costs of mistakenly treating one-shot encounters as repeated encounters. By experience, subjects can be persuaded that their default assumptions are in error; however, it remains an empirical question as to how much, and under what conditions, such defaults continue to influence decisions. Interpreting the design of emotion mechanisms in this light suggests both new experiments to tease apart the cues involved as well as new theory exploring the evolution of strategies in an environment with stochastically varying group sizes.

Emotions as Mechanisms That Manipulate Time Horizons

Aggression and punishment as strategies which change the behavior of other individuals rely upon a fundamental logic: Reactions to current transgressions must be sufficiently costly to the target to deter future transgressions. However, deterrence is costly. It is costly for one individual to inflict harm on another, and these costs must be paid in the present even though the benefits will be reaped in the future. This leads to a puzzle because humans, like virtually all other animals studied, steeply discount the future. Anger may effectively solve this problem, motivating people to respond to transgressions and overriding the tendency to discount the future (Daly and Wilson 1988; Lerner and Keltner 2001; Fessler and Haley, this volume). In fact, anger sometimes seems to be disproportionate to the magnitude of the transgression, perhaps because the anger system sums the costs of prospective future transgressions and then substitutes this sum for the actual cost of the present transgression (Frank 1988). Reputational effects may magnify emotional responses because the payoffs of deterrence are multiplied when third parties observe the response or hear others gossip about the response. Thus, anger may be expressed even in one-shot interactions if reputational effects are important (Nisbett and Cohen 1996).

CONCLUSION

A number of problems remain unsolved for understanding cooperation outside kin selection. In this report, we have summarized the group discussion of cognitive and emotional mechanisms which instantiate possible solutions. This discussion has certainly not been exhaustive. Several important topics remain unexplored. Many mechanisms which were selected by inclusive fitness may have been later exapted (i.e., put to a new purpose) to serve roles in nonkin cooperation, and we have neglected phylogeny in almost every aspect of the discussion. Theory of mind and the attribution of intentions is a large and important topic in cognition and cooperation, which we have only touched upon here. Our discussion of justified defections in indirect reciprocity invokes intentionality and suggests that individuals use attributed intentions in guiding their cooperative behavior, and strategies in the iterated PD such as Contrite Tit-for-Tat (Boyd 1989) necessarily invoke the communication of intentions.

Many of the experiments and studies we have discussed, especially with respect to human friendship, are inadequate to address many of the newer questions. With respect to human friendship, this is because the studies in social psychology were conducted with different questions in mind. Thus a number of new experiments and observations will be needed to address the concerns raised in this report. We have tried to suggest such empirical investigations where obvious, but we think that inventive experimenters and fieldworkers will see many more, just as ingenious theoreticians will no doubt see many promising modeling possibilities that we have missed.

REFERENCES

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

Axelrod, R. 1984. The Evolution of Cooperation. New York: Basic.

Axelrod, R., and W.D. Hamilton. 1981. The evolution of cooperation in biological systems. *Science* **211**:1390–1396.

Barr, A. 2001. Social dilemmas, shame-based sanctions, and shamelessness: Experimental results from Rural Zimbabwe. In: CSAE Working Paper WPS/2001.11. Oxford: Centre for the Study of African Economics.

Barrett, L., and S.P. Henzi. 2001. The utility of grooming in baboon groups. In: Economics in Nature, ed. R. Noë, J.A.R.A.M. van Hooff, and P. Hammerstein, pp. 119–145. Cambridge: Cambridge Univ. Press.

Bechara, A., A.R. Damasio, H. Damasio, and S. Anderson. 1994. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* **50**:7–15.

Bechara, A., H. Damasio, and A. Damasio. 2000. Emotion, decision making, and the orbitofrontal cortex. *Cerebral Cortex* **10**:295–307.

Bechara, A., H. Damasio, D. Tranel, and A.R. Damasio. 1997. Deciding advantageously before knowing the advantageous strategy. *Science* **275**:1293–1295.

Bendor, J., R.M. Kramer, and S. Stout. 1991. When in doubt ...: Cooperation in a noisy prisoner's dilemma. *J. Conflict Resol.* **35**:691–719.

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

Boyd, R. 1989. Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *J. Theor. Biol.* **136**:47–56.

Boyd, R., and P.J. Richerson. 1988. The evolution of reciprocity in sizeable groups. *J. Theor. Biol.* **132**:337–356.

Boyd, R., and P.J. Richerson. 1989. The evolution of indirect reciprocity. *Soc. Netw.* **11**:213–236.

Buss, D. 2001. Cognitive biases and emotional wisdom in the evolution of conflict between the sexes. *Curr. Dir. Psychol. Sci.* **10**:219–223.

Chang, A. 2002. The Relationship between Recalling a Relevant Past Experience and Vigilance for Cheaters and Altruists. B.Sc., Dept. of Psychology, McMaster Univ., Hamilton, Ontario, Canada.

Chase, I.D. 1982. Dynamics of hierarchy formation: The sequential development of dominance relationships. *Behaviour* **80**:218–240.

Chase, I.D., C. Tovey, D. Spangler-Martin, and M. Manfredonia. 2002. Individual differences versus social dynamics in the formation of animal dominance hierarchies. *Proc. Natl. Acad. Sci. USA* **99**:5744–5749.

Clutton-Brock, T.H., and G.A. Parker. 1995. Punishment in animal societies. *Nature* **373**:209–216.

Cords, M. 2002. Friendship among adult female blue monkeys. *Behaviour* **139**:291–314.

Cosmides, L. 1989. The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition* **31**:187–276.

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

Damasio, A.R. 1994. Descartes' Error: Emotion, Reason, and the Human Brain. New York: Grosset and Putnam.

de Waal, F.B.M. 1997. The chimpanzee's service economy: Food for grooming. *Evol. Hum. Behav.* **18**:375–386.

Dugatkin, L.A. 1997. Cooperation among Animals: An Evolutionary Perspective. Oxford: Oxford Univ. Press.

Enquist, M., and O. Leimar. 1993. The evolution of cooperation in mobile organisms. *Anim. Behav.* **45**:747–757.

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

Farrell, J., and M. Rabin. 1996. Cheap talk. *J. Econ. Persp.* **10**:103–118.

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., G. Kirchsteiger, and A. Riedl. 1993. Does fairness prevent market clearing?: An experimental investigation. *Qtly. J. Econ.* **108**:437–460.

Fessler, D.M.T. 1999. Toward an understanding of the universality of second-order emotions. In: Beyond Nature or Nurture: Biocultural Approaches to the Emotions, ed. A. Hinton, pp. 75–116. New York: Cambridge Univ. Press.

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

Frank, R. 1988. Passions within Reason: The Strategic Role of the Emotions. New York: Norton.

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.

Hamilton, W.D. 1964. The genetical evolution of social behavior. II. *J. Theor. Biol.* **7**:17–52.

Hart, B.L., and L.A. Hart. 1992. Reciprocal allogrooming in impala, *Aepyceros melampus*. *Anim. Behav.* **44**:1073–1083.

Haselton, M.G., and D.M. Buss. 2000. Error management theory: A new perspective on biases in cross-sex mind reading. *J. Pers. Soc. Psych.* **78**:81–91.

Hemelrijk, C.K. 1994. Support for being groomed in long-tailed macaques, *Macaca fasicularis*. *Anim. Behav.* **48**:479–481.

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.

Henzi, S.P., and L. Barrett. 2002. Infants as a commodity in a baboon market. *Anim. Behav.* **64**:915–921.

Hey, J. 1997. Mitochondrial and nuclear genes present conflicting portraits of human origins. *Mol. Biol. Evol.* **14**:166–172.

Johnstone, R.A. 2001. Eavesdropping and animal conflict. *Proc. Natl. Acad. Sci. USA* **98**:9177–9180.

Leimar, O., and P. Hammerstein. 2001. Evolution of cooperation through indirect reciprocity. *Proc. Roy. Soc. Lond. B* **268**:745–753.

Lerner, J., and D. Keltner. 2001. Fear, anger, and risk. *J. Pers. Soc. Psych.* **81**:146–159.

Loewenstein, G. 1996. Out of control: Visceral influences on behavior. *Org. Behav. Hum. Dec. Proc.* **65**:272–292.

Maynard Smith, J. 1991. Honest signalling: The Philip Sidney game. *Anim. Behav.* **42**:1034–1035.

Maynard Smith, J. 1994. Must reliable signals always be costly? *Anim. Behav.* **47**:1115–1120.

Milinski, M. 1987. TIT FOR TAT in sticklebacks and the evolution of cooperation. *Nature* **325**:433–435.

Milinski, M., J.H. Lühti, R. Eggler, and G.A. Parker. 1997. Cooperation under predation risk: Experiments on costs and benefits. *Proc. Roy. Soc. Lond. B* **264**:1239–1247.

Milinski, M., D. Pfluger, D. Külling, and R. Kettler. 1990. Do sticklebacks cooperate repeatedly in reciprocal pairs? *Behav. Ecol. Sociobiol.* **27**:17–21.

Milinski, M., D. Semmann, T.C.M. Bakker, and H.-J. Krambeck. 2001. Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proc. Roy. Soc. Lond. B* **268**:2495–2501.

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

Milinski, M., and C. Wedekind. 1998. Working memory constrains human cooperation in the prisoner's dilemma. *Proc. Natl. Acad. Sci. USA* **95**:13,755–13,758.

Nesse, R.M. 2001. The smoke detector principle: Natural selection and the regulation of defenses. In: Unity of Knowledge: The Convergence of Natural and Human Science, ed. A.R. Damasio, A. Harrington, J. Kagan et al., vol. 935, pp. 75–85. New York: New York Academy of Sciences.

Nisbett, R.E., and D. Cohen. 1996. Culture of Honor: The Psychology of Violence in the South. Boulder, CO: Westview Press.

Nowak, M.A., and K. Sigmund. 1992. Tit for tat in heterogenous populations. *Nature* **355**:250–253.

Nowak, M.A., and K. Sigmund. 1993. A strategy of win-stay, lose-shift outperforms tit for tat. *Nature* **364**:56–58.

Nowak, M.A., and K. Sigmund. 1998a. The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**:561–574.

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

Packer, C., D.A. Gilbert, A.E. Pusey, and S.J. O'Brien. 1991. A molecular genetic analysis of kinship and cooperation in African lions. *Nature* **351**:562–565.

Relethford, J.H. 1998. Genetics of modern human origins and diversity. *Ann. Rev. Anthro.* **27**:1–23.

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.

Schino, G. 2001. Grooming, competition, and social rank among female primates: A meta-analysis. *Anim. Behav.* **62**:265–271.

Seyfarth, R.M., and D.L. Cheney. 1984. Grooming, alliances, and reciprocal altruism in vervet monkeys. *Nature* **308**:541–543.

Silk, J.B., E. Kaldor, and R. Boyd. 2000. Cheap talk when interests conflict. *Anim. Behav.* **59**:423–432.

Stopka, P., and R. Graciasova. 2001. Conditional allogrooming in the herb-field mouse. *Behav. Ecol.* **12**:584–589.

Sugden, R. 1986. The Economics of Rights, Co-operation, and Welfare. New York: Blackwell.

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

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

Wason, P.C. 1968. Reasoning about a rule. *Qtly. J. Exp. Psych.* **20**:273–289.

Wedekind, C., and V.A. Braithwaite. 2002. The long-term benefits of human generosity in indirect reciprocity. *Curr. Biol.* **12**:1012–1015.

Wedekind, C., and M. Milinski. 2000. Cooperation through image scoring in humans. *Science* **288**:850–852.

Williams, G.C., and R.M. Nesse. 1991. The dawn of Darwinian medicine. *Qtly. Rev. Biol.* **66**:1–22.

Wolfpoff, M. 1998. Concocting a divisive theory. *Evol. Anthro.* **7**:1–3.