In P.M. Kappeler and C.P. van Schaik (eds) Cooperation in Primates and Humans: Mechanisms and Evolution. Springer-Verlag: Berlin, 2006

Chapter 4

Reciprocal altruism: 30 years later

Robert Trivers

"Two are better than one; because they have a good reward for their labour. For if they fall, the one will lift up the other: but woe to him that is alone when he falleth; for he hath not another to help him up. Again, if two lie together, then they have heat: but how can one be warm alone? And if one prevails against him, two shall withstand him; and a three-fold cord is not easily broken." (Ecclesiastes 4, 9–12; King James Version).

4.1 Introduction

A little over 30 years ago, I had the good fortune of publishing my first scientific paper on reciprocal altruism, a subject that had not yet been addressed from an evolutionary standpoint. Hamilton's (1964) great work on kinship and altruism made it clear that in humans there existed a major form of altruism that could not be explained by kinship. Its elaboration was responsible for the complex economic systems in which we now live and its regulation could plausibly be explained by a system of interconnected human emotions, including feelings of friendship, gratitude, sympathy, guilt, moralistic aggression, a sense of justice and (I would now add) forgiveness.

I brought no great talents to this enterprise, beyond a willingness to take the evolutionary problem seriously and to model evolutionary logic on easily inferred psychological facts regarding our own behavior (for a description of how the paper was written, see Trivers 2002). The paper was certainly timely. My 600 reprints were quickly exhausted and the evolutionary idea was off and running. There now exists a very large literature on the subject and many subareas have advanced far beyond my original paper.

The purpose of the present paper is to provide a personal review of some major developments since my paper. These include the Prisoner's Dilemma (PD) as a model for reciprocal altruism, other models and third-party observer effects. I concentrate on the human sense of justice and the selective forces likely to have molded it. In the process, I discuss recent empirical work (using economic games) that bears on our sense of fairness and what seems to me the most plausible way to interpret these results. I neglect many important topics, for example, discrimination against cheaters in symbioses (see Sachs et al. 2004).

4.2
What was accomplished in 1971

It is well to pause for a moment and remember what the pre-evolutionary period looked like. Although Charles Darwin and George Williams had devoted a sentence or two to the subject of reciprocal altruism, both anthropology and social psychology were thoroughly pre-Darwinian in their thinking. Anthropology recognized the importance of reciprocity but not the problem of the cheater within the system. Conceptual confusion was illustrated by the effort to define parental investment as an example of 'reciprocity between the generations'. You invest in your children and they in theirs. Social psychology saw that 'prosocial' tendencies were important in life but failed to see any problem in how they evolved (indeed, did not even raise the question) and therefore failed to differentiate between obvious subcategories such as kin-directed versus reciprocal altruism. Evolutionary theory had nothing to offer beyond group selection and general inattention.

In my paper, elementary distinctions were emphasized. Altruism is suffering a cost to confer a benefit. Reciprocal altruism is the exchange of such acts between individuals so as to produce a net benefit on both sides. Reciprocal altruism is one kind of return-benefit altruism. There can be a variety of ways in which an act of altruism can initiate a causal chain leading to a return benefit to the actor, of which reciprocal altruism is but one example. A warning call which keeps group members alive may bring an immediate return benefit and eating your cleaner may bring a cost in lost interaction with the now dead cleaner. I was primarily interested in reciprocal altruism and would probably have skipped these two examples of return-benefit altruism if I had any nonhuman examples of reciprocal altruism. At the same time, the cleaning symbiosis example demonstrated altruism without kinship and the bird example provided a host of return-effect alternatives to a kinship explanation. In any event, the larger area of return-benefit altruism has become much more important, although it is not always conceptualized this way. Thus, group selection language is often formally equivalent to the language of return effects, though this may be obscured or denied. A good recent review, especially in the context of symbioses, can be found in Sachs et al. (2004).

A second distinction of some importance concerns the cheater or non-reciprocator in the system. I deliberately chose the term cheater, even though the neutral term non-reciprocator (later, defector) was more precise, because the emotive and intuitive powers of 'cheater' were attractive to me. The key point is to figure out how the cheater may harm itself, often by the counteraction of others. For me, the simplest was simply to break off the relationship, thereby cutting one's own losses and coincidentally reducing the benefits of cheating. The second is direct punishment, but since this is itself costly, I imagined that it would evolve after mere non-reciprocation. I called it 'moralistic aggression' deliberately, since it had a moral flavor but was not necessarily moral.

I naturally wanted the domain of my paper to be as large as possible, so I avoided, wherever possible, any limiting assumptions. Both Darwin (1871) and Williams (1966) quickly restricted reciprocal altruism to species with complex

cognitive powers, while I preferred to imagine that once reciprocity got underway, the requisite cognitive powers would quickly evolve. More recently, possible cognitive limitations to the evolution of reciprocity have been emphasized (Stephens & Hauser 2004).

Finally, the paper sketched out a few obvious ways in which the analysis could be extended to a complex, multi-party system, with norms, infractions of norms, observer effects, collective punishment, and so on. These are topics that have become very active (and, in some cases, contentious) areas of research in recent years.

4.3
Tit-for-tat

My approach began with reality and attempted to model its evolution. In particular, I turned to everyday life for an account of what seemed to me the key emotions regulating the system and I dressed these up with data from social psychology. This can be a very effective first step, and indeed a continuing source of ideas for theoretical development, but one soon wants to have theory that begins with fundamental assumptions and generates different possible worlds. The key first advance was provided by Axelrod & Hamilton (1981). Building on the results of Axelrod's computer tournaments to see who could devise the best strategy for playing iterated games of PD (Rapaport's tit-for-tat [TfT] was the winner), they were able to show that TfT is evolutionarily stable as long as there is sufficient probability of future interaction (always-defect is also stable). TfT was the simplest strategy introduced in the original tournaments and has only two rules: begin cooperative and then on the next move do whatever your partner did on the previous one (i.e. cooperate or defect).

The simplicity is itself beguiling. It immediately widens the range of situations in which to look for reciprocal relations. Why not bacteria? They certainly are capable of a contingent strategy with neighbors in which an individual produces a (cooperative) chemical first and then produces whichever chemical the neighbor does (cooperative or selfish). There is no question that a rich world of social interactions will be found in bacteria, sometimes unifying the behavior of millions or billions of separate bacteria, but disentangling kinship, green beard and reciprocal effects will not be easy. A defector mutant in a sea of cooperators will initially be unrelated to neighbors at the key locus, and *vice versa*. Exciting empirical work is now emerging on the PD and cooperation in viruses (Turner & Chao 1999), bacteria (Velicer & Yu 2003) and yeast (Greig & Travisano 2004).

The success of the TfT strategy also gives new meaning to the value in life of imitation or, at least, responding in kind. Imitation is classically viewed as a valuable form of learning but it may also often be the appropriate response in reciprocal relations. Do unto others as they have just done unto you. If someone is nice, be nice; if not so nice, not so nice; if nasty, be nasty, and so on. But there are limitations to the value of this rule, as seen in the large costs associated with outbreaks of reciprocal spite (see below).

TfT and Axelrod & Hamilton (1981) spread like wildfire through the behavioral literature, so that soon all reciprocal interactions seemed to be formulated in terms of the PD. The effect was to distance the analysis from subjective experience to the point where I had to translate papers back into English in order to understand what was being said; for example, the excellent work on predator inspection in sticklebacks (Milinksi 1987, Milinski et al. 1990). It soon seemed that theorists and empiricists alike were forgetting that iterated games of PD amount to a highly artificial model of social interactions; each successive interaction simultaneous, costs and benefits never varying, options limited to only two moves, no errors, no escalated punishment, no population variability within traits and so on. In fact, almost all of these simplifying assumptions have now been shown to introduce important effects.

4.4
Beyond tit-for-tat in the Prisoner's Dilemma

In a brilliant series of papers, Nowak, Sigmund and colleagues explored the consequences of relaxing the assumptions built into the original game of iterated PDs. In the process, they outlined a plausible account of the way in which new strategies may displace old ones in an evolving two-dimensional social world where the two dimensions are the probabilities that an individual will cooperate in response to either the partner's cooperation or defection on the previous move. Nowak (1990) first showed that the introduction of errors into the system brought forth the value of some degree of forgiveness by the tit-for-tatter and some relaxation in the rigidity of always doing what your partner just did. In a simple iterated PD between two tit-for-tatters, a mistaken move by one on any move (say, defect on the first) will put the two actors permanently out of synchrony with each other, one cooperating, the other defecting, then *vice versa*, and so on, never achieving a relationship of strong reciprocal benefits. Various strategies of partial forgiveness can cut through this dilemma and outcompete TfT. The first to be described was 'generous tit for tat' (gen TfT) in which a modified tit-for-tatter responds always to cooperation with cooperation but responds to defection some of the time with cooperation; in effect, forgiving the partner for that defection (Nowak & Sigmund 1992). If the partner is a tit-for-tatter (generous or original), cooperation will have been re-established, the error corrected. The expected frequency of forgiving is determined by the precise payoff matrix of the game (one-third in the original payoff conditions of Axelrod's tournament).

This work invites us to consider the many mistakes that can be made in life, not just accidental breakdowns in the underlying causal machinery. We can try to second-guess the situation and guess wrong, we can be in a negative state because of other interactions which spill over into this situation; there are 'innocent bystanders' (Nowak & Sigmund 1992) and so on. Error is intrinsic to life but perhaps more frequent with some kinds of strategies or in some kinds of settings than in others, an avenue of investigation that may prove fruitful.

Nowak & Sigmund (1993) then showed that yet another strategy did very well in a world of tit-for-tatters, generous tit-for-tatters and full-time cooperators (all-c). Called 'win-stay, lose-shift', it merely repeats the same move as on the previous if it was rewarded and changes its move if it was punished. Put another way, if your partner just cooperated, keep doing what you are doing; if the partner just defected, switch. This strategy has three benefits: (i) it protects against exploitation, (ii) corrects errors and (iii) exploits a naive cooperator. Note that it cannot invade in a world of defectors, because it switches every other interaction to cooperate, where it is exploited. But if TfT and gen TfT have driven out all-defect (all-d), then all-c will spread by drift, inviting the success of win-stay, lose-shift.

Nowak & Sigmund (1994) have also analyzed the alternating PD where, instead of acting simultaneously, players alternate roles of donor and recipient. This would seem to mimic real life more closely, though including some (non-zero) probability of repeating the same role would do even better. Generous TfT does relatively better than win-stay, lose-shift and the latter can only work well in the alternating game if memory extends back more than one move. When students play iterated PD games that are either simultaneous or alternating, they are more successful in the simultaneous game with win-stay, lose-shift strategies but, as expected, more successful in the alternating game with gen TfT (Wedekind & Milinksi 1996). In addition, the win-stay, lose-shift people adjusted their strategy to protect better against all-d and exploit better all-c.

The greater cognitive complexity of win-stay, lose-shift has also been nicely confirmed by Milinski & Wedekind (1998). They asked students to play the iterated PD either continuously or interrupted after each round by a memory task (playing the game 'Memory') that acts to reduce working memory capacity. Most people playing the regular game end up adopting win-stay, lose-shift strategies, while the remaining play gen TfT. With the memory task imposed, people play the cognitively simpler gen TfT relatively more often. In addition, students who continued to use complex win-stay, lose-shift strategies were successful doing so but became less good at playing the Memory game.

There have been additional developments along these lines, e.g. modeling the effect of the spatial structure of the social neighborhood in selecting for cooperation (Nowak & May 1992) or imagining different levels of satisfaction for win-stay, lose-shift strategies (Posch et al. 1999). More generally, it has been argued that most two-party games should be modeled as a series of interactions in which partners negotiate the final outcome (McNamara et al. 1999).

4.5
Expanding beyond the Prisoner's Dilemma

Important findings have appeared which evade the limitations of the PD (and my cursory treatment of the area is purely a function of my ignorance). Introducing individual variation alone tends to drive out the inflexible strategies we have reviewed above (Johnson et al. 2002). Cooperation can thrive through the appearance of a strategy that begins small and then, when matched, raises in-

vestment (Roberts & Sherratt 1998). Another way to make reciprocity more viable, is simply to break down a large transfer into a series of contingent small ones. If as a hermaphroditic sea bass I approach you laden with eggs, and give you all to fertilize, I have nothing left with which to bargain for the privilege of fertilizing your eggs (sperms being less expensive than eggs). You are free to swim away and seek sperm elsewhere. But if I offer you a few and then no more unless you offer a few in return, we can spend a happy afternoon in reciprocal egg exchanges, which eventually add up to what we could have achieved in one exchange (Fischer 1988). Similar behavior may have evolved in polychaete worms (Sella et al. 1997). And in impalas, mutual grooming is broken down into a series of short bouts which can be terminated on evidence of non-reciprocation (Hart & Hart 1992). Certainly in human life, we recognize the principle as well, much more willing to make a loan or extend some help, if it can be broken down into smaller parts, with evidence of some positive effect available from earlier dispensations before the later ones are extended. A general theory of when parceling is expected (and how) would be most welcome.

Recently, a new game has been introduced as a metaphor for cooperation, the snowdrift game. This is the same as 'hawk vs. dove' in aggressive encounters. Two cars are stuck in a snowdrift. Each driver can free its car by removing the drift, which also frees the other car, or the two can work together (and the relative fraction of effort can be made to vary). The key is that cooperative behavior produces direct benefits to another individual and to self simultaneously. Analysis of this model produces striking departures from findings in the PD. For example, in the PD, spatial structure often favors cooperation, but in the snowdrift, the opposite is true (Hauert & Doebeli 2004). Spatial structure reduces the frequency of cooperation for a wide range of parameters. This is because the payoff is greatest if one's strategy differs from that of one's partner or neighbor. Hauert & Doebeli (2004) show that (depending on assumptions regarding payoffs) many situations modeled already as PD games may better be modeled as snowdrift games. In the continuous snowdrift game in which a cost is incurred to transfer a benefit to self and other, a uniform population often gives way to a bifurcated one, with a stable proportion of individuals making large (cooperative) investments coexisting with a set of defectors who invest little (Doebeli et al. 2004).

4.6 Possible green beard effects

When I first tried to model reciprocal altruism, I gave each individual in a two-party interaction a different locus guiding their reciprocal tendencies. I did this to avoid a so-called green-beard effect (alleles at a locus able, in effect, to recognize themselves in another individual, even if the rest of the genome is uncorrelated between the two: Hamilton 1964) but I quickly became convinced (with Bill Hamilton's encouragement) that such complexities were better left to the future. Yet, they may be important. While kinship can be excluded as a factor in some cases, this is not so easy for green beard effects. Consider TfT models

which conventionally assume genes at a single (haploid) locus, in which case, by definition, when two individuals hook up, they aid themselves but also identical copies at the same locus in another individual. Thus, by its very definition, reciprocal altruism, similarity between partners is expected, perhaps at the loci directing the altruism.

A start in this direction has been made by Riolo et al. (2001a) who claim to have shown that cooperation without reciprocity can evolve via tag-based altruism where a tag is any observable phenotypic trait (including behavioral) toward whom tag-bearers direct their altruism. The problem with such systems is that they are vulnerable to the evolution of individuals who display the tag but not the altruism, leading to the rapid collapse of tag-based cooperation (Roberts & Sherratt 2001). It is clear that any successful tag-based model requires a series of additional assumptions whose plausibility has not been addressed (Riolo et al. 2001b). For another attempt to model greenbeard effects in human reciprocal altruism, see Price (in press).

An interesting example of green beard cooperation has been described in some detail for the side-blotched lizard, *Uta stansburiana* (Sinervo & Clobert 2003). Three color morphs occur in males. The blue morph is cooperative and two such males will settle close to each other, without regard to relatedness, and cooperate in driving off yellow males who attempt to sneak copulations within the territories of blue males. The orange morph, in turn, dominates blue males. Although individuals assort themselves non-randomly by color, it remains unclear if this also truly occurs at an underlying switch gene controlling color and functionally-related behavioral characters. Another oft-cited example of greenbeard effects was described in fire ants but has not been confirmed in subsequent work (Keller & Ross 1998, KG Ross pers. com.). The only clear examples of genetic green-beard effects operate at the level of cell-cell interactions, as in gametophytic factors in plants, cell-adhesion molecules in cellular slime molds and, probably, in mother-fetal interactions in mammals (reviewed by Burt & Trivers 2006).

4.7 Intrapersonal reciprocity?

Let us briefly consider another genetic hypothesis. It is possible to imagine internal genetic conflict within an individual that may, in part, be ameliorated by reciprocal effects within the genome. This is a completely novel perspective that remains to be confirmed (or disconfirmed) but the logic is fairly simple to describe. The outbred, heterozygous genetic system can be conceived of as an example of reciprocal cooperation between (usually) unrelated genomic halves (maternal and paternal). A kinship-based system would show high internal relatedness (i.e. inbreeding between parents) but with the costly side-effect of increased homozygosity. Although the two halves are typically unrelated, no conflict is expected most of the time, because the fates of the two genomes are tightly bound together in the survival of the same individual. But at reproduction, the genes are separated and sent into new individuals and this gives rise to multiple

avenues for conflict in the germline as genes are selected to increase their own replication at the expense of the larger organism (a massive topic, reviewed by Burt & Trivers 2006). Axelrod & Hamilton (1981) imagined that even here reciprocal relations could restrain selfish tendencies, if, for example, a chromosome in one cell could respond to evidence of transmission ratio distortion by its homolog in another cell by doing the same. This would sometimes lead to both chromosomes driving, producing a trisomic individual. In principle, this argument could partly explain the dramatic increase in trisomy 21 in humans associated with maternal age.

In humans, there is one striking exception to the rule that internal genetic conflict is limited to conflict over genetic transmission. Genomic imprinting permits parent-specific gene expression, so that maternal and paternal genes can act according to their exact degrees of relatedness to each parent (and relatives related through them; reviewed by Haig 2002). This conflict is now well established for early life in mice and humans, paternally-active genes (with the maternal copy silent) tending to extract more resources from the mother and *vice versa* for maternally-active ones. And there are strong hints of conflict later in life as well (e.g. effects on adult behavior and brain structure).

An interesting possibility is that within-individual reciprocal relations may develop between the maternal and paternal halves of our genomes (Haig 2003). There are many situations in which the two could benefit by foregoing the costs of opposing each other and agreeing on some fair bargain. Haig (2003) wonders whether oppositely-imprinted genes might, over time, evolve complex strategies of cooperation, defection and punishment that are conditional upon the expression of oppositely-imprinted genes. It would be very interesting if such genes could create stable paternal and maternal personalities that interacted as individuals. If so, interactions with parental molding would be expected, since one set of genes in the offspring, associated with one personality, has interests exactly aligned with one parent but not the other, and *vice versa*. Does the degree of reciprocity between a couple affect the degree of reciprocity within their children?

4.8 A false negative

Every living theory generates a backlash of some sort, often, new theoretical work claiming to undermine or sharply limit the original argument. Sometimes this leads to a lively exchange, which helps to clarify matters, as in the case of my theory of parent-offspring conflict (Trivers 1974), countered immediately by Alexander (1974; logic and evidence reviewed by Godfray 1995a). My paper on reciprocal altruism generated no such initial attack. Indeed, the paper was followed a decade later by Axelrod & Hamilton (1981), which only served to strengthen the underlying theory and point the way toward many future developments. It fell to Boyd & Richerson (1988) to publish the first serious challenge, not to the logic itself but to the notion that it could work in any but small groups (5 individuals is a large group, 25 very large). The paper is now often cited as

a reason to skip quickly beyond reciprocal altruism (and the iterated PD) to a group-beneficial view of our sense of fairness (e.g. McElreath et al. 2003). Fehr (2004) goes the extra step and tells us that reciprocal altruism can only evolve in groups of two, a depressing fact, were it true.

Boyd & Richerson (1988) claim that their "model closely resembles models of reciprocity in pairs" and from it, one can safely conclude that "the conditions necessary for the evolution of reciprocity become extremely restrictive as group size increases". In fact, their model does not model reciprocity within pairs and their pessimism is entirely a result of artificial assumptions at the outset, primarily that "If an individual switches from defection to cooperation every other member of the group is better off". This strange assumption is justified because it "formalizes the idea that cooperation benefits other members of the group". But why insist that cooperation between two must benefit all? If I groom you (or feed you or protect you), why does every one else have to benefit? Boyd & Richerson (1988) are pretending to model dyadic relations but have forced on such interactions a large group effect, which inevitably grows in power with group size. They have created a world in which defectors automatically benefit from the trading of favors between reciprocators. This is an unnatural assumption.

Put differently, Boyd and Richerson seem to think that when we move to n-person groups, the natural extension of 2-person model is to impose the same kind of payoff structure on the entire n-person group. Indeed, a two-person interaction within an n-person group can not, by assumption, occur because it automatically affects all other members of the n-party group. This is highly unrealistic. In fact, the opposite is more likely, that most interactions within the group continue to be dyadic or triadic. Most interaction in real time is a sequence of fast decisions about how to behave toward a small number of individuals who are immediately present. What group size of n does is to affect whom one is likely to be interacting repeatedly with. It is a mistake to collapse all these interactions into one large n-adic interaction with the additional restriction near-universal benefit or cost for any act.

Elementary logic suggests that there can be no dramatic limitation to reciprocal altruism as groups include more than a few. Even in groups of 40, a tit-for-tatter should be able to learn in 40 costly interactions who are the fellow cooperators, hence capping its losses early, say after a week or a month, with months and perhaps years of benefits to flow from successful cooperation. Recent mathematical modeling, using the PD in finite populations, shows that TfT invades at very low frequencies in moderate-sized ($N \sim 30$) groups (Nowak et al. 2004). We now turn to the way in which reciprocal altruism can be extended to n-party interactions within such groups.

4.9
Indirect reciprocity, image scoring and reputation

What were just a few sentences in my original paper, the possibility of three-party interactions with important observer effects, has in the past 10 years become an entire sub-discipline. Although Alexander (1987) appreciated the importance of the subject, the key step was to model the evolution of 'image scoring', the tendency to assign an individual a positive image or a negative image, depending on whether that individual was seen to act altruistically or selfishly toward a third party (Nowak and Sigmund 1998a,b). Positive images induce altruism from others and negative images selfishness. Since in this model a strategy was also included of being indifferent to the images, it was possible to show that image-scoring itself could evolve. Observer effects, in turn, are strong enough to induce cooperation and fairness, along with punishment of non-cooperators (Sigmund, Hauert and Nowak 2001). This entire subject has now been beautifully reviewed by Sigmund and Nowak (*in press*), with numerous novel insights. A subtlety of some importance is that an actor who is seen to fail to give to another can be so acting because the actor is stingy or because the recipient is unworthy (i.e. itself stingy in interactions with others). Likewise, punishing an unworthy individual may improve your image, while punishing a worthy one will have the opposite effect. But how is the observer to have this kind of detailed knowledge? Opportunities for deception would seem to be rife, increasing furthe the cognitive complexity of reciprocal altruism, with associated selection on mental traits.

Once again, a pleasing feature of the explicit theory is that it has been coupled to laboratory experiments – games played for money – that explore the underlying dynamics with real people. For example, under experimental conditions of anonymity, in which individuals can respond to whether other have been generous or not, donations go more frequently to those who themselves have been more generous in earlier interactions (Wedekind and Milinski 2000). Also, while generosity induced generosity, it did not produce a net benefit, but this turned out to be an effect of how long the game was played (n=6). When the game is twice as long, a significant positive effect emerges (Wedekind & Braithwaite 2002). The work shows that cooperation through image scoring does occur even when individuals are known only by an arbitrary marker. It provides an easy way out of measuring the net effect, which turns out to be more positive the more rounds that are played. For additional work along these lines, consult chapter 14 (see also Milinski et al 2001, 2002).

4.10
A sense of justice

Perhaps the most important implication of my paper (for me, at least) was that it laid the foundation for understanding how a sense of justice evolved. At the time, the sense of justice in humans was usually considered a product of culture and upbringing with no biological component. I thought that grounding a sense of justice in biology would only strengthen our attachment to it and naively assumed that those with a self-professed interest in justice would greet the work warmly. This turned out to be true for the great philosopher of the subject, John Rawls (1971), while the pseudo-radicals of the 1970s tore after my work (and sociobiology more generally) like so many rabid dogs after a fleeing rabbit, missing the point entirely (Trivers 1981).

I was surprised on rereading my paper recently to find no reference to justice itself but only to the weaker term 'norms'. I apparently did not get around to using the proper term until Trivers (1981, 1985). The argument, in any case, is the same. Even in two-party interactions, but especially in multi-party ones, one needs a standard by which to judge deviations from symmetrical (or fair) interactions, the better to detect cheaters. Such cheating is expected to generate strong emotional reactions because unfair arrangements, repeated often, may exact a very strong cost in inclusive fitness. In that sense, an attachment to fairness or justice is self-interested and we repeatedly see in life, as expected, that victims of injustice feel the pain more strongly than do disinterested bystanders and far more than do the perpetrators. This is not to say that we have no response to injustice visited on people very far from us in space (or even time) but this does not imply that our sense of justice evolved with these distant events in mind.

The first appearance of the concern in children is thoroughly self-interested ('but that's not fair, mommy') and those who would push a group selection (or mere cultural diffusion) view of the sense of justice would do well to ask themselves when last they heard a child say, 'mommy, daddy, I am acting unfairly, please stop me'. Self-interest is often confused with selfishness, a mistake given wide prominence by Dawkins' (1976) misuse of the word 'selfish'. 'Enlightened self-interest', after all, is meant to call attention to the value for our own selves of funneling certain benefits through others.

It is easy (in the United States, at least) to underestimate the power of our sense of justice. For example, the neoconservative architects of the current U.S. bloodletting in Iraq are fond of saying that the United States is disliked in other countries not because its policies are perceived as unjust (and certainly not because they are unjust) but because the U.S. is envied for its size, power and wealth. Envy, however, is a trivial emotion compared to our sense of injustice. To give one possible example, you do not tie explosives to yourself to kill others because you are envious of what they have, but you may do so if these others and their behavior represent an injustice being visited on you and yours day after day and year after year, often with little alternative action available to you, as in Palestine, where the people suffer robbery of land, water and life and are denied little more than rifles in self-defense, while their next-door Israeli neighbors (and occupiers) are armed with nuclear weapons and the latest in U.S. lethal technology.

There is growing evidence that something like a sense of fairness has evolved in a range of non-human primates who, in turn, practice reciprocal altruism (see de Waal & Brosnan, this volume; for chimp reciprocity in the wild, see Watts 2002). Thus, there appears to be an aversion to inequity in both capuchins and chimpanzees (Brosnan & de Waal 2003, Brosnan et al. 2004). In chimpanzees, ungenerous individuals are attacked when they beg for food, as are those who do

not support those who just supported them (de Waal 1992b). These observations are consistent with the view that a sense of fairness evolved slowly over a long period of time, first in dyadic relationships and then more widely. The number of dyadic relationships should not be underestimated; there may be a fair split between the interests of parent and offspring, sibling and sibling, near-neighbor and near-neighbor, and so on.

4.11
The experimental study of reciprocal altruism

One of the more welcome developments in the study of reciprocity, and the human sense of justice, is the development of economic games that attempt to mimic real life situations but can be played in the lab for real money. These games can be played cross-culturally and altered in a series of ways to explore relevant causal factors. This is a considerable advance over the world of social psychology, with its reliance on paper-and-pencil tests of human dispositions, artificial situations difficult to interpret and (sometimes) deception of the subjects under study. Indeed, Rapaport & Chammah (1965) saw this very clearly when they argued not only for the theoretical utility of the PD but its value in generating empirical data to test the theory. As we have seen, experiments in which people play iterated games of PD have provided valuable evidence on the cost-benefit ratio of such strategies as win-stay, lost-shift or gen TfT.

There is now a very large and excellent literature on economic games (see Roth 1995, Smith 2003). Many have been played in the laboratory under controlled conditions, games with names like the ultimatum game, dictator game, public goods game and so on. They can be played single shot or iterated, anonymous or non-anonymous, with and without onlookers, and they can be analyzed theoretically with simple models.

One game will serve as an example, the single-shot, anonymous ultimatum game. The game is played once by a proposer and a responder (Güth et al. 1982, Burnham 2002). The proposer is given an amount of money (by the experimenter) to split with an unknown responder. He (or she) proposes a split and if the responder accepts the offer, the two split the money accordingly, but if player two rejects the offer, neither player receives any money. No further interactions occur between the two. In the standard economic model of maximizing financial gain, responders are expected to take whatever they are offered, as long as it is not zero, and, thus, proposers are expected to make very low offers and to try to keep most of the money. But this is not what research shows (Forsythe et al. 1994). Offer modes and medians are 40-50%, offer means are 30-40%, and offers below 20% are usually rejected (Camerer 2003). This result now has been replicated, with some interesting variation, in scores of studies, with varying stakes, around the world (Henrich et al. 2005).

The problem arises in how to interpret these results. It is generally agreed that they show our attachment to a sense of fairness, even when this costs us money but why do we act this way? To many evolutionists, including myself, this sense of justice or fairness benefits us in everyday life by protecting us from unfair arrangements that harm our inclusive fitness. We are expected to react negatively to unfair offers by others, not out of envy of their extra portion, but because they chose to inflict this unfair offer on us and the unchallenged repetition of such behavior is expected in the future to inflict further costs on our inclusive fitness. Consistent with this, people accept lower offers from a computer than they do from another person, even though both offers offer identical payoffs to two humans (Blount 1995). This approach assumes that our responses were never selected to perform in the highly unusual, one-shot, anonymous interaction in a lab, with payoffs underwritten by a third party. Put another way, these experimental results seem on their face neither unexpected nor puzzling.

By contrast, some social scientists now playing these games have opted for a very different view (e.g. Fehr & Fischbacher 2003, Gintis et al. 2003). According to them, the results prove that our sense of fairness cannot have a self-interested function, all possibility of return effects having been removed. Instead, it must have been selected to benefit the group or appeared by some process of cultural diffusion. This they call 'strong reciprocity', to differentiate it from the 'weak' reciprocity of classic reciprocal altruism

4.12
'Strong reciprocity?'

In defense of this interpretation of the ultimatum game, Bowles & Gintis (2003) tell us "We do not think that subjects are unaware of the one-shot setting, or unable to leave their real-world experiences behind with repeated interactions at the laboratory door". Surely, awareness is irrelevant. You can be aware that you are in a movie theatre watching a perfectly harmless horror film and still be scared to death. As for leaving real-world experiences at the laboratory door, I know of no species, humans included, that leaves any part of its biology at the laboratory door; not prior experiences, nor natural proclivities, nor ongoing physiology, nor arms and legs, nor whatever. That is the whole point of experimental work. You bring living creatures into the lab (ideally, whole) to explore causal factors underlying their biology, the mechanisms in action. You do not imagine that you have thereby solved the problem of evolutionary origin; that is, that you can shortcut the problem of evolutionary function by simply assuming that the organism's actions in the lab represent evolved adaptations to the lab.

People do not leave their religion at the laboratory door and this alone can induce 'observer effects' for those who imagine that God is scrutinizing their every action (D. Stahl, pers. com.). Indeed, the impression that God will punish malefactors may have been promoted to encourage cooperation (Johnson & Krüger 2004). Nor do people leave their culture at the door, and it has been claimed cross-culturally that the higher the degree of market integration and cooperation in daily life, the more prosocial individuals act in anonymous games in the lab (Henrich et al. 2005). Nor do people leave their testosterone at the lab door and unpublished work shows that such levels are positively associated with rejection rates in men (T. Burnham unpubl. data). Apparently, those who promote 'strong reciprocity' believe that a little verbiage ('you know not who the other

actor is nor will you ever know – and *vice versa*') is sufficient to put the human being into a state of suspended animation such that he or she automatically and appropriately adjusts to the full evolutionary implications of this novel, experimental situation.

This argumentation is supported, they argue, by the fact that humans do make discriminations in everyday life based on chance of future interaction, diminishing cooperation as frequency of interaction decreases (exactly as expected according to the theory of reciprocal altruism). But the fact that y, a function of x, is decreasing as x approaches zero, does not mean that the function passes through zero simultaneously on both axes. How do they handle this deficiency? With a little jargon, "individuals should be fully capable of taking a 'zero-baseline' of cooperation" (Fehr & Henrich 2003). This merely assumes what needs to be shown and is unlikely on its face.

These authors (and others) also claim that human evolution was characterized by frequent one-shot encounters with important fitness effects (Fehr & Henrich 2003). Leaving aside murder, this seems most unlikely, especially when one considers that what is meant is one-shot anonymous interactions of the form of an ultimatum game. As I have pointed out elsewhere, social interactions are intrinsically repeat interactions, certainly over very small time scales of seconds and minutes but almost always over longer time periods as well (Trivers 2004). What they are calling one-shot encounters were really repeat interactions lasting at least minutes and hours, if not days and months. Imagine I am walking through the woods with a man I have never met before and am unlikely to see after the walk, with not a soul in sight. He points to five low-hanging apples and suggests that he stand on my shoulders to pick them, after which he will give me one. I am likely to argue the point immediately (and a chance to communicate even once prior to an anonymous ultimatum game has been shown to affect the interaction; Rankin 2003). Even if I agree with him, as he lands on the ground, I may knock three apples out of his hand and run, or strangle him, taking apples and any other property he has (after all, observer and other return effects have been ruled out). This interaction all takes place over a span of several minutes. Thus, even an imaginary 'one-shot' encounter really consists of a series of interactions.

I have a close relative who, among other virtues, has a well-developed sense of spite. She has a long memory for slights and will repay in kind. As I described the ultimatum game to her, with her share of the pie (as recipient) now reduced to 30%, her fingernails began (literally) to dig into the table and her face contorted with anger (looks just like her everyday self to me, I thought). The intensity was remarkable. Those fingers were set to dig into someone's face, if need be. When Jamaican youngsters are given low offers, they often respond with a flash of anger and something like, 'this is all I get?'. Anger is not a mere emotion, it is (costly) physiological arousal for immediate aggressive action. Anger makes no sense without the possibility of future interaction. Neurophysiological work on the ultimatum game is also consistent with biological arousal for future interaction. Functional Magnetic Resonance Imaging (fMRI) work shows that unfair offers are met with activation of part of the recipient's brain (the anterior cingulate) involved in negative emotions, primarily anger and disgust, and

control functions involving conflict, and the higher the activation, the greater the chance that such an offer will be rejected (Sanfey et al. 2003). In short, in our everyday behavior and neurophysiology we respond to so-called one-shot encounters as if they were the first in a chain of interactions.

The errors I have drawn attention to are but a few of many. An individual who turns down an 80:20 offer hardly benefits the group; 100% of resources disappear at once (Burnham & Johnson 2005). Indeed, moral philosophers have shown how the group-benefit approach to justice is inferior to 'justice as fairness', in which a fair arrangement is defined not by whether it maximizes group output, but by whether individuals would accept it if they did not know in advance which position they occupied (Rawls 1971). For those seeking a more in-depth treatment of the 'strong reciprocity' approach and its failings, see Burnham & Johnson (2005). These include idiosyncratic use of language. A lifetime of trading benefits with others in a fair manner is apparently 'weak reciprocity' while 'strong reciprocity' is a single, take-it-or-leave-it interaction with no possible reciprocity, underwritten by a third party. Likewise, punishment of unfair behavior is called 'altruistic punishment' on the assumption that such behavior has a net cost for the actor while benefiting others, something that needs to be shown, not assumed.

Finally, for me there is a feeling in all of this of déjà vu, all over again. There has been a long history in the social sciences of assuming whatever is necessary in order to make the argument that is desired. Is it necessary to assume that there is little or no genetic variation in human social traits in order to push an extreme environmental interpretation? So be it. Can we assume that humans are rational utility maximizers, where utility is anything the organism wishes to see maximized, and that on this foundation we can safely build social theory? Be our guest. May we assume that people's behavior in the highly unusual one-shot ultimatum game reflects how they were selected to act in precisely this situation? So let it be granted. On the bright side, there is actually some progress here. The first position denies both genetics and evolution. The second merely assumes that these subjects are irrelevant, while the third embraces evolutionary logic in principle but promptly gives it a truncated form.

4.13 Forgiveness and revenge

Forgiveness may play an important role in a system of reciprocal altruism, especially where the latter has degenerated into a system of reciprocal spite. Forgiveness tends to short-cut the spite, potentially saving enormous amounts of energy, both outward-directed and inner-consumed. There is strong evidence that positive emotions are associated with high immune performance (Rosenkranz et al. 2003) and that manipulations of mood which increase positive affect, e.g. through meditation, are themselves associated with an increase in immune response (Davidson et al. 2003). There is no direct evidence for similar benefits from forgiveness, but certainly many who have suffered grievous loss through the action of others have spoken of the corrosive internal effects of retaining the

spiteful mentation of hatred and revenge (Cole 2004). As we saw earlier, the existence of mistakes automatically selects for a degree of forgiveness in everyday life. Forgiveness may also be efficient from a cognitive perspective, dropping a contradiction in one's mind between past losses, present failure to redress the injustice, which then requires planning for future activity. Letting go is letting go of a burdensome scheme of mentation. Set against this, is the impulse toward revenge and retribution. This seems most likely in social arrangements that endure over long periods of time, including several generations, so that long-delayed revenge is both possible and possibly instrumental in protecting later-borne relatives. A careful treatment of the evolutionary dynamics of forgiveness and/or revenge would be most welcome.

4.14 Justice and truth

Elementary logic suggests a possible connection in individuals between apprehension of the truth and commitment to justice in social relations. An immoral stance often requires deception, with its inevitable effects on self-deception (Trivers 2000). This tendency may be more pronounced at the top of a social hierarchy, where truth-telling by others is emphasized as a necessary virtue, while an illusion of the same by self is promulgated. There is a natural inclination to over-emphasize the justice of one's own position and thus to under-emphasize that of one's opponent, thereby underestimating the strength of their resistance. This asymmetry is especially striking in aggressive invasions of the land of others, where there ought to be a natural (moral and physical) presumption in favor of the prior occupants.

4.15 Torture

The demonstration that altruistic punishment (or moralistic aggression) is pleasurable to the punisher, as judged neurophysiologically (de Quervain et al. 2004), invites the speculation that this has made the appearance of torture more likely. Torture is the laser-like application of highly spiteful activities toward inducing pain, fear, shame, madness, and so on, in others. It may provide pleasure for the torturers, especially if they believe such action is morally justified, as in the words of U.S. Vice President Cheney, that the 'worst of the worst' are incarcerated in Guantanamo, Cuba and in Iraq. Recently, we have learned in the United States that internal government documents argue that torture is not torture if there is no organ failure or imminent death, a redefinition that eliminates nearly all non-fatal forms of torture. Such are the achievements of self-deception in the service of moralistic aggression.

4.16 Summary

Since 1971, an enormous and very sophisticated literature has grown up on reciprocal altruism and cooperation more generally, both in biology and economics. Noteworthy has been the success of generating a detailed series of theoretical findings, both within and outside the Prisoner's Dilemma paradigm. In the latter, we see conditions under which a series of strategies can displace each other or survive together: all-d, TfT, gen TfT, all-c and win-stay, lose-shift. We have seen alternatives to the PD such as the snowdrift game, which may produce very different effects (e.g. of spatial structure). And we have considered two genetic hypotheses of interest, green-beard altruism and within-individual genetic reciprocity, both of which await confirmation. We have outlined the way in which reciprocal altruism may generate a sense of justice as a means of guarding against cheaters and we have explored an alternative interpretation, popular in some circles, that our sense of fairness evolved without regard to return effects, a view for which there is as yet no useful, positive evidence. All of these areas of research are unusually vibrant and productive at the present time and much valuable new work is published literally every month. This work shows every promise of uniting, at last, important parts of economics with evolutionary biology.

Acknowledgments

I am grateful to Jim Bull, Terry Burnham, Peter Godfrey-Smith, Christoph Hauert, Marc Hauser, Peter Kappeler, Martin Nowak, Carel van Schaik, Dale Stahl, Claus Wedekind, Richard Wrangham and Darine Zaatari for many, helpful comments. I am grateful to Terry Burnham for first alerting me to the difficulty of interpreting the results of economic games.

References

Alexander, R.D. 1974. The evolution of social behavior. Annual Review Ecology Systematics 5: 325-383.

Axelrod, R. and Hamilton, W.D. 1981. The evolution of cooperation. Science 211: 1390-1396.

Blount, S. 1995. When social outcomes aren't fair: the effect of causal attributions on preferences. Organizational Behavior and Human Decision Processes 63: 131-144.

Bowles, S. and Gintis, H. 2003. Origins of human cooperation. In P. Hammerstein (ed) Genetic and Cultural Evolution of Cooperation. Cambridge, MA: MIT, pp 429-443.

Boyd, R. and Richerson, P.J. 1988. The evolution of reciprocity in sizable groups. Journal Theoretical Biology 132: 337-356.

Brosnan, S.F. and de Waal, F.B.M. 2003. Monkeys reject unequal pay. Nature 425: 297-299.

Brosnan, S.F., Schiff, H.C. and de Waal, F.B.M. 2004. Tolerance for inequity may increase with social closeness in chimpanzees. Proceedings Royal Society London B 272: 253-258

Burnham, T.C. 2002. Ultimatum Games. In: The Encyclopedia of Cognitive Science. London: Nature Publishing Group, pp 238-245.

Burnham, T.C., and Johnson, D. 2005. The biological and evolutionary logic of human cooperation. Analyse and Kritik 27 (2)

Burnham, T.C. manuscript. Caveman economics: proximate and ultimate causes of non-materially maximizing behavior.

Burt, A. and Trivers, R. 2005. Genes in Conflict: The Biology of Selfish Genetic Elements. Cambridge, MA: Harvard University Press.

Camerer, C.F. 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton: Princeton University Press.

Cole, E. 2004. Bone to Pick: Of Forgiveness, Reconciliation, Reparation and Revenge. New York: Atria Books.

Crozier, R. and Pamilo, P. 1996. Evolution of Social Insect Colonies: Sex Allocation and Kin Selection. New York: Oxford University Press.

Davidson, R.J., Kabat-Zinn, J., Schumacher, J., Rosenkranz, M., Muller, D., Santorelli, S.F., Urbanowski, F., Harrington, A., Bonus, K., and Sheridan, J.F. 2003. Alterations in brain and immune function produced by mindfulness meditation. Psychosomatic Medicine 65: 564-570.

de Quervain, D.J.-F, Fischbacher, U., Treyer, V., Scellhammer, M., Schnyder, U., Buck, A., and Fehr, E. 2004. The neural basis of altruistic punishment. Science 305: 1254-1258.

de Waal, F.B.M. 1992. The chimpanzees sense of social regularity and its relation to the human sense of justice. In R.D. masters and M. Gruter (Eds.) The Sense of Justice: Biological Foundations of Law. Newbury Park: Sage, pp 241-255.

Doebeli, M., Hauert, C. and Killingback, T. 2004. The evolutionary origin of cooperators and defectors. Science 306: 859-862.

Feh, C. 1999. Alliances and reproductive success in Camargue stallions. Animal Behaviour 57: 705-713.

Fehr, E. 2004. Don't lose your reputation. Nature 432: 449-450.

Fehr, E. and Fischbacher , U. 2003. The nature of human altruism. Nature 425: 785-791.

Fehr, E. and Henrich, J. 2003. Is srong reciprocity a maladaption? On the evolutionary foundations of human altruism. In P. Hammerstein (ed). Genetic and Cultural Evolution of Cooperation. Cambridge, MA: MIT Press, pp 455-82.

Fischer EA. 1988. Simultaneous hermaphroditism, Tit-for-Tat, and the evolutionary stability of social systems. Ethology and Sociobiology: 9:119-36.

Forsythe, R., Horowitz, J., Savin, N.E. and Sefton, M. 1994. Fairness in simple bargaining experiments. Games and Economic Behavior 6: 347-369.

Gintis, H., Bowlses, S., Boyd, R. and Fehr, E. 2003. Explaining altruistic behavior in humans. Evolution and Human Behavior 24: 153-172.

Godfray, H. 1995. Evolutionary theory of parent-offspring conflict. Nature 376: 133-138.

Greig, D. and Travisano, M. 2004. The prisoner's dilemma and polymorphism in yeast and *SUC* genes. Proceedings Royal Society London B (Supplement) 271: S25-S26.

Güth, W., Schmittberger, Schwartze, B. 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3: 367-388.

Haig, D. 2002. Kinship and Genomic Imprinting. New Brunswick: Rutgers University Press.

Haig, D. 2003. On intrapersonal reciprocity. Evolution Human Behavior 24: 418-425.

Hart, B.L. and Hart, L.A. 1992. Reciprocal allgrooming in impala, *Aepyceros meleanopus*. Animal Behaviour 44: 1073-1083.

Hauert, C. and Doebeli, M. 2004. Spatial structure often inhibits the evolution of cooperation in the snowdrift game. Nature 428: 643-646.

Hauser, M.D., Chen, M.K., Chen, F., Chang, E., Chuang, E. 2003. Give unto others: genetically unrelated cotton-top tamarin monkey preferentially give food to those who altruistically give food back. Proceedings Royal Society London B 270: 2363-2370.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. 2001. In Search of Homo economicus: Behavioral experiments in 15 small-scale societies. American Economic Review 91: 73-78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N.S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F.W., Patton, J.Q., and Tracer, D. 2005. 'Economic man' in cross-cultural perspective: behavioral experiments in 15 small-scale societies. Behavioral and Brain Sciences

Johnson, D.P., Stopka, P. and Bell, J. 2002. Individual variation evades the Prisoner's Dilemma. BMC Evolutionary Biology 2(15):1-8.

Johnson, D.D.P. and Krüger, 2004. The good of wrath: supernatural punishment and the evolution of cooperation. Political Theology 5: 159-176.

Linklatter, W.L. and Cameron, E.Z. 2000, Tests for cooperative behaviour between stallions. Animal Behaviour 60: 731-743.

McElreath, R., Clutton-Brock, T.H., Fehr, E., Fessler, D.M.T., Hagen, E.H., Hammerstein, P., Kosfeld, M., Millinski, M., Silk, J.B. Tooby, J., Wilson, M.I.. 2003. Group report: the role of cognition and emotion in cooperation. In P. Hammerstein (ed). Genetic and Cultural Evolution of Cooperation. Cambridge, MA: MIT Press, pp 125-152.

McNamara, J.M., Gasson, C.E. and Houston, A.I. 1999. Incorporating rules for responding into evolutionary games. Nature 401: 368-371.

Milinski, M. 1987. TIT FOR TAT in sticklebacks and the evolution of cooperation. Nature 325: 433-437.

Milinski, M., Külling, D., and Kettler, R. 1990. Tit for Tat: stickleback (*Gastereus aculeatus*) "trusting" a cooperative partner. Behavioral Ecology 1: 7-11.

Nowak, M. 1990. Stochastic strategies in the prisoner's dilemma. Theroetical Population Biology 38: 93-112.

Nowak, M, and Sigmund, K. 1992. Tit for tat in heterogeneous populations. Nature 355: 250-253.

Nowak, M. and Sigmund, K. 1993. A strategy of win-stay, lose-shift outperforms tit-for-tat in the Prisoner's Dilemma game. Nature 364: 56-58.

Nowak, M. and Sigmund, K. 1994. The alternating prisoner's dilemma. Journal of theoretical Biology 168: 219-226.

Nowak, M.A. and May, R.M. 2004. Evolutionary games and spatial chaos. Nature 359: 826-829.

Nowak, M.A., Sasaki, A., Taylor, C. and Fudenberg, D. 2004. Emergence of cooperation and evolutionary stability in finite populations. Nature 428: 646-649.

Posch, M., Pichler, A. and Sigmund, K. 1999. The efficiency of adapting aspiration levels. Proceedings Royal Society London B 266: 1427-1435.

Price, M.E. in press. Monitoring, reputation and "greenbeard" reciprocity in a Shuar work team. Journal of Organizational Behavior

Rankin, F.R. 2003. Communication in ultimatum games. Economic Letters 81: 267-271.

Rapaport, A. and Chammah, A.M. 1965. Prisoner's Dilemma. Ann Arbor: University of Michigan Press.

Rawls, J. 1971. A Theory of Justice. Cambridge,MA: Harvard University Press.

Riolo, R., Cohen, M.D. and Axelrod, R. 2001a. Evolution of cooperation without reciprocity. Nature 414: 441-443.

Riolo, R., Cohen, M.D. and Axelrod, R. 2001b. Does similarity breed cooperation: Riolo et al reply. Nature 418: 500.

Roberts, G. and Sherratt, T.N. 2002. Does similarity breed cooperation? Nature 418: 499-500.

Rosenkranz, M.A., Jackson, D.C., Dalton, K.M., Dolski, I., Ryff, C.D., Singer, B.H., Muller, D., Kalin, N.H. and Davidson, R.J. 2003. Affective style and *in vivo* immune response: neurobehavioral mechanisms. Proceedings National Academy of Sciences (U.S.A.) 100: 1114-11152.

Roth, A.E. 1995. Bargaining experiments. In: J.H.Kagel and A.E. Roth (Eds), Handbook of Experimental Economics. Princeton, NJ: Princeton University Press.

Sachs, J., Mueller, U.G., Wilcox, T.P. and Bull, J.J. 2004. The evolution of cooperation. Quarterly Review Biology 79: 135-160.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E. and Cohen, J.D. 2003. The neural basis of economic decision-making in the ultimatum game. Science 300: 1755-1758.

Sella G, Premoli MC, Turri F. 1997. Egg trading in the simultaneously hermaphroditic polychaete worm *Ophryotrocha gracilis* (Huth). Behavioral Ecology 8:83-86.

Sethi, R., and Somanathan, E. 2003. Understanding reciprocity. Journal of Economic Behavior and Organization 50: 1-27.

Sinervo, B., and Clobert, J. 2003. Morphs, dispersal behavior, genetic similarity, and the evolution of cooperation. Science 300: 1949-1951.

Smith, V. 2003. Constructivist and ecological rationality in economics. American Economic Review 93: 465-508.

Stevens, J.P. and Hauser, M.D. 2004. Why be nice? Psychological constraints on the evolution of cooperation. Trends in Cognitive Science 8: 60-65.

Trivers, R.L. 1971. The evolution of reciprocal altruism. Quarterly Review Biology 46:25-57.

Trivers, R.L. 1974. Parent-offspring conflict. American Zoologist 14: 249-264.

Trivers, R.L. and Hare, H. 1976. Haplodiploidy and the evolution of the social insects. Science 30: 253-269.

Trivers, R. 1981. Sociobiology and politics. In E. White (Ed.) Sociobiology and Human Politics. Lexington: Lexington Books, pp 1-43.

Trivers, R. 1985. Social Evolution. Menlo Park: Benjamin/Cummings.

Trivers, R. 2000. The elements of a scientific theory of self-deception.. Annals N.Y. Academy Sciences 907: 114-131.

Trivers, R. 2002. Natural Selection and Social Theory: Selected Papers of Robert Trivers. New York: Oxford University Press.

Trivers, R. 2004. Mutual benefits at all levels of life. Science 304: 965-965.

Turner, P.E. and Chao, L. 1999. Prisoner's dilemma in an RNA virus. Nature 398:441-443.

Velicer, G.C. and Yu, Y-t.N. 2003. Evolution of novel cooperative swarming in the bacterium *Myxococcus xanthus*. Nature 425: 75-78.

Watts, D.P. 2002. Reciprocity and interchange in the social relationships of wild male chimpanzees. Behaviour 139: 343-370.

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

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