CHAPTER 42

Explaining altruistic behaviour in humans

Herbert Gintis, Samuel Bowles, Robert Boyd and Ernst Fehr

42.1. Introduction

The explanatory power of inclusive-fitness theory and reciprocal altruism (Hamilton, 1964; Williams, 1966; Trivers, 1971) convinced a generation of researchers that what appears to be altruism—personal sacrifice on behalf of others—is really just long-run self-interest. Richard Dawkins, for instance, struck a responsive chord when, in *The Selfish Gene* (1979, p. vii), he confidently asserted: "We are survival machines—robot vehicles blindly programmed to preserve the selfish molecules known as genes. ... This gene selfishness will usually give rise to selfishness in individual behaviour." Dawkins allows for morality in social life, but it must be socially imposed on a fundamentally selfish agent. "Let us try to teach generosity and altruism," he advises, "because we are born selfish" (Dawkins, *The selfish Gene*, p. 3). Yet even social morality, according to R. D. Alexander, the most influential ethicist working in the Williams– Hamilton tradition, can only superficially transcend selfishness. In *The Biology of Moral Systems* (1987), Alexander asserts (p. 3): "ethics, morality, human conduct, and the human psyche are to be understood only if societies are seen as collections of individuals seeking their own self-interest." In a similar state of explanatory euphoria, Ghiselin (1974) claims (p. 247): "No hint of genuine charity ameliorates our vision of society, once sentimentalism has been laid aside. What passes for cooperation turns out to be a mixture of opportunism and exploitation ... Scratch an altruist, and watch a hypocrite bleed."

However, recent experimental research has revealed forms of human behaviour involving interaction among unrelated individuals that cannot be explained in terms of self-regarding preferences. One such trait, which we call strong reciprocity (Gintis, 2000b; Henrich *et al.*, 2001), is *a predisposition to cooperate with others, and to punish those who violate the norms of cooperation, at personal cost, even when it is implausible to expect that these costs will be repaid either by others or at a later date.*

In this chapter, we present evidence supporting strong reciprocity. We then explain why, under conditions plausibly characteristic of the early stages of human evolution, a small fraction of strong reciprocators could invade a population of self-regarding types, and why strong reciprocity is an evolutionarily stable strategy. Throughout this chapter, we use the term 'self-regarding' rather than the more common term 'self-interested' to avoid the (uninteresting, we believe) question as to whether it is selfish to help others if that is how one 'maximizes utility'. Although most of the evidence we report is based on behavioural experiments, the same behaviours are regularly observed in everyday life, and of great relevance for social policy (Gintis *et al.*, 2005).

Despite the fact that strong reciprocity is altruistic, our results do not contradict traditional

evolutionary theory. A gene that promotes self-sacrifice will die out unless those who are helped carry the mutant gene, or its spread is otherwise promoted. In a population without structured social interactions of individuals, behaviours of the type found in our experiments and illustrated in our models could not have evolved. However, multi-level selection and gene–culture coevolutionary models support cooperative behaviour among non-kin (Feldman *et al.*, 1985; Sober and Wilson, 1998; Gintis, 2000b, 2003; Henrich and Boyd, 2001; Bowles *et al.*, 2003). These models, some of which are discussed below, are not vulnerable to the classic critiques of group selection by Williams (1966), Dawkins (1976), Maynard Smith (1976), Rogers (1990), and others.

An alternative account of strong reciprocity is that in our hunter-gatherer ancestral environment, strong reciprocity was not altruistic, but rather was individually fitness-maximizing, as it allowed individuals to develop a reputation for being both willing to cooperate, yet committed to retaliating against those who betray their trust. In the contemporary environment, so the argument goes, strongly reciprocal behaviour persists in situations where it is altruistic, but these situations would rarely have arisen in our hunter-gatherer past, where anonymous, one-shot interactions were supposedly extremely rare. We think this alternative is unlikely, and address the issue in Section 42.7. Indeed, we argue in Section 42.6 that through gene–culture coevolution, our species developed a whole range of social emotions, including shame, guilt, pride and honour, that both promoted individual well-being and a high level of social cooperation. The remainder of the chapter is devoted to a deeper analysis of social emotions.

42.2. Experimental evidence: strong reciprocity in the labour market

Strong reciprocity is most clearly exhibited in laboratory experiments. In one such experiment (Fehr *et al.*, 1997) the experimenters divided a group of 141 subjects (college students who had agreed to participate in order to earn money) into a set of 'employers' and a larger set of 'employees'. The rules of the game are as follows. If an employer hires an employee who provides effort e and receives a wage w, the employer's pay-off p is 100 times the effort e, minus the wage w that he must pay the employee ($p = 100e - w$), where the wage is between zero and 100 ($0 = w = 100$), and the effort between 0.1 and 1 ($0.1 = e = 1$). The pay-off u to the employee is then the wage he receives, minus a 'cost of effort', $c(e)$ ($u = w - c(e)$). The cost-of-effort schedule $c(e)$ is constructed by the experimenters such that supplying effort $e = 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0 costs the employee $c(e) = 0$, 1, 2, 4, 6, 8, 10, 12, 15 and 18, respectively. All pay-offs are converted into real money that the subjects are paid at the end of the experimental session.

The sequence of actions is as follows. The employer first offers a 'contract' specifying a wage w and a desired amount of effort e^*. A contract is made with the first employee who agrees to these terms. An employer can make a contract (w, e^*) with at most one employee. The employee who agrees to these terms receives the wage w and supplies an effort level e, which *need not equal the contracted effort*, e^*. In effect, there is no penalty if the employee does not keep his promise, so the employee can choose any effort level, $e \in [0.1, 1]$, with impunity. Although subjects may play this game several times with different partners, each employer–employee interaction is a one-shot (non-repeated) event. Moreover, the identity of the interacting partners is never revealed. This experiment is especially relevant because it models a situation that could have resulted in one-shot interactions among acquaintances in small-scale societies. An individual makes a promise, and then, because monitoring is not possible, fails to keep it.

If employees are self-regarding, they will choose the zero-cost effort level, $e = 0.1$, no matter what wage is offered them. Knowing this, employers will never pay more than the minimum necessary to get the employee to accept a contract, which is 1 (assuming only integral wage offers are permitted). The employee will accept this offer, and will set $e = 0.1$. Since $c(0.1) = 0$, the employee's pay-off is $u = 1$. The employer's pay-off is $p = 0.1 \times 100 - 1 = 9$.

In fact, however, this self-regarding outcome rarely occurred in this experiment. The average

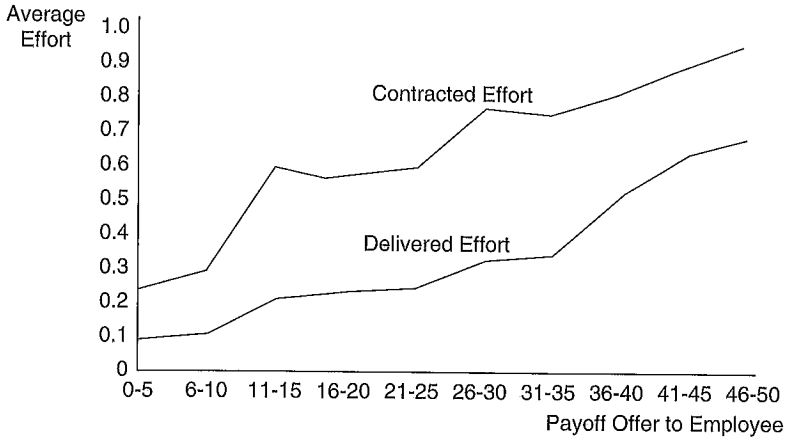

Fig. 42.1 Relation of contracted and delivered effort to worker pay-off (141 subjects). From Fehr et al. (1997).

net pay-off to employees was $u = 35$, and the more generous the employer's wage offer to the employee, the higher the effort provided. In effect, employers presumed the strong reciprocity predispositions of the employees, making quite generous wage offers and receiving higher effort, as a means to increase both their own and the employee's pay-off, as shown in Figure 1. Similar results have been observed in Fehr *et al.* (1993, 1998).

Figure 40.1 also shows that, though most employees are strong reciprocators, at any wage rate there still is a significant gap between the amount of effort agreed upon and the amount actually delivered. This is not because there are a few 'bad apples' among the set of employees, but because only 26% of employees delivered the level of effort they promised! We conclude that strong reciprocators are inclined to compromise their morality to some extent, just as we might expect from daily experience.

The above evidence is compatible with the notion that the employers are purely self-regarding, since their beneficent behaviour *vis-á-vis* their employees was effective in increasing employer profits. To see if employers are also strong reciprocators, following this round of experiments, the authors extended the game by allowing the employers to respond reciprocally to the actual effort choices of their workers. At a cost of 1, an employer could increase or decrease his employee's pay-off by 2.5. If employers were self-regarding, they would of course do neither, since they would not interact with the same worker a second time. However, 68% of the time, employers punished employees who did not fulfil their contracts, and 70% of the time, employers rewarded employees who overfulfilled their contracts. Indeed, employers rewarded 41% of employees who exactly fulfilled their contracts. Moreover, employees expected this behaviour on the part of their employers, as shown by the fact that their effort levels increased significantly when their bosses gained the power to punish and reward them. Underfulfilling contracts dropped from 83% to 26% of the exchanges, and overfulfilled contracts rose from 3% to 38% of the total. Finally, allowing employers to reward and punish led to a 40% increase in the net pay-offs to all subjects, even when the pay-off reductions resulting from employer punishment of employees are taken into account. Several researchers have predicted this general behaviour on the basis of general real-life social observation and field studies, including Homans (1961), Blau (1964) and Akerlof (1982). The laboratory results show that this behaviour has a motivational basis in strong reciprocity and not simply long-term material self-interest.

We conclude from this study that the subjects who assume the role of 'employee' conform to internalized standards of reciprocity, even when they know that there are no material repercussions from behaving in a self-regarding manner. Moreover, subjects who assume the role of

'employer' expect this behaviour and are rewarded for acting accordingly. Finally, 'employers' draw upon the internalized norm of rewarding good and punishing bad behaviour when they are permitted to punish, and 'employees' expect this behaviour and adjust their own effort levels accordingly.

42.3. Experimental evidence: the ultimatum game

In the ultimatum game, under conditions of anonymity, two players are shown a sum of money, say $10. One of the players, called the 'proposer,' is instructed to offer any number of dollars, from $1 to $10, to the second player, who is called the 'responder'. The proposer can make only one offer. The responder, again under conditions of anonymity, can either accept or reject this offer. If the responder accepts the offer, the money is shared accordingly. If the responder rejects the offer, both players receive nothing.

Since the game is played only once and the players do not know each other's identity, a self-regarding responder will accept any positive amount of money. Knowing this, a self-regarding proposer will offer the minimum possible amount, $1, and this will be accepted. However, when actually played, *the self-regarding outcome is never attained and never even approximated.* In fact, as many replications of this experiment have documented, under varying conditions and with varying amounts of money, proposers routinely offer respondents very substantial amounts (50% of the total generally being the modal offer), and respondents frequently reject offers below 30% (Güth and Tietz, 1990; Roth *et al.*, 1991; Camerer and Thaler, 1995).

The ultimatum game has been played around the world, but mostly with university students. We find a great deal of individual variability. For instance, in all of the above experiments a significant fraction of subjects (about a quarter, typically) behave in a self-regarding manner. But, among student subjects, average performance is strikingly uniform from country to country.

To expand the diversity of cultural and economic circumstances of experimental subjects, Henrich *et al.* (2004) undertook large cross-cultural study of behaviour in various games including the ultimatum game and the public goods game. Twelve experienced field researchers, working in 12 countries on four continents, recruited subjects from 15 small-scale societies exhibiting a wide variety of economic and cultural conditions. These societies consisted of three foraging groups (the Hadza of East Africa, the Au and Gnau of Papua New Guinea, and the Lamalera of Indonesia), six slash-and-burn horticulturists (the Aché, Machiguenga, Quichua, and Achuar of South America, and the Tsimané and Orma of East Africa), four nomadic herding groups (the Turguud, Mongols, and Kazakhs of Central Asia, and the Sangu of East Africa) and two sedentary, small-scale agricultural societies (the Mapuche of South America and Zimbabwe farmers in Africa). We can summarize our results as follows.

1. The canonical model of self-regarding behaviour is not supported in any society studied. In the ultimatum game, for example, in all societies either respondents, or proposers, or both, behaved in a reciprocal manner.
2. There is considerably more behavioural variability across groups than had been found in previous cross-cultural research. While mean ultimatum game offers in experiments with student subjects are typically between 43% and 48%, the mean offers from proposers in our sample ranged from 26% to 58%. While modal ultimatum game offers are consistently 50% among university students, sample modes with these data ranged from 15% to 50%. In some groups rejections were extremely rare, even in the presence of very low offers, while in others, rejection rates were substantial, including frequent rejections of hyper-fair offers (i.e. offers above 50%). By contrast, the most common behaviour for the Machiguenga was to offer zero. The mean offer was 22%. The Aché and Tsimané distributions resemble American distributions, but with very low rejection rates. The Orma and Huinca (non-Mapuche Chileans living among the Mapuche) have modal offers near the centre of the distribution.
3. Differences among societies in 'market integration' and 'cooperation in production' explain a substantial portion of the behavioural

variation between groups: the higher the degree of market integration and the higher the pay-offs to cooperation, the greater the level of cooperation and sharing in experimental games. The societies were rank-ordered in five categories: 'market integration' (how often do people buy and sell, or work for a wage), 'cooperation in production' (is production collective or individual), plus 'anonymity' (how prevalent are anonymous roles and transactions), 'privacy' (how easily can people keep their activities secret), and 'complexity' (how much centralized decision-making occurs above the level of the household). Using statistical regression analysis, only the first two characteristics, market integration and cooperation in production, were significant, and they together accounted for 66% of the variation among societies in mean ultimatum game offers.

4. Individual-level economic and demographic variables did not explain behaviour either within or across groups.
5. The nature and degree of cooperation and punishment in the experiments was generally consistent with economic patterns of everyday life in these societies. In a number of cases the parallels between experimental game play and the structure of daily life were quite striking.

Nor was this relationship lost on the subjects themselves. Here are some examples.

- The Orma immediately recognized that the public goods game was similar to the harambee, a locally initiated contribution that households make when a community decides to construct a road or school. They dubbed the experiment 'the harambee game' and gave generously (mean 58% with 25% maximal contributors).
- Among the Au and Gnau, many proposers offered more than half the pie, and many of these 'hyper-fair' offers were rejected! This reflects the Melanesian culture of status-seeking through gift-giving. Making a large gift is a bid for social dominance in everyday life in these societies, and rejecting the gift is a rejection of being subordinate.
- Among the whale hunting Lamalera, 63% of the proposers in the ultimatum game divided the pie equally, and most of those who did not offered more than 50% (the mean offer was 57%). In real life, a large catch, always the product of cooperation among many individual whalers, is meticulously divided into pre-designated parts and carefully distributed among the members of the community.
- Among the Aché, 79% of proposers offered either 40% or 50%, and 16% offered more than 50%, with no rejected offers. In daily life, the Aché regularly share meat, which is distributed equally among all other households, irrespective of which hunter made the kill. The Hadza, unlike the Aché, made low offers and had high rejection rates in the ultimatum game. This reflects the tendency of these small-scale foragers to share meat, but with a high level of conflict and frequent attempts of hunters to hide their catch from the group.
- Both the Machiguenga and Tsimané made low ultimatum game offers, and there were virtually no rejections. These groups exhibit little cooperation, exchange or sharing beyond the family unit. Ethnographically, both show little fear of social sanctions and care little about 'public opinion'.
- The Mapuche's social relations are characterized by mutual suspicion, envy, and fear of being envied. This pattern is consistent with the Mapuche's post-game interviews in the ultimatum game. Mapuche proposers rarely claimed that their offers were influenced by fairness, but rather by a fear of rejection. Even proposers who made hyper-fair offers claimed that they feared rare spiteful responders, who would be willing to reject even 50/50 offers.

42.4. **Experimental evidence: the public goods game**

The public goods game has been analysed in a series of papers by the social psychologist Toshio Yamagishi (1986a,b, 1988), by the political scientist Elinor Ostrom *et al.* (1992), and by economist Ernst Fehr and his coworkers (Gächter and Fehr, 1999; Fehr and Gächter, 2000, 2002). These researchers uniformly found that, although they rarely attained efficiency, *groups exhibit a much higher rate of cooperation than can be expected assuming the standard economic model of the self-interested actor*, and this is especially

the case when subjects are given the option of incurring a cost to themselves in order to punish freeriders.

A typical public goods game consists of a number of rounds, say 10. The subjects are told the total number of rounds, as well as all other aspects of the game. The subjects are paid their winnings in real money at the end of the session. In each round, each subject is grouped with several other subjects, for example three others, under conditions of strict anonymity. Each subject is then given a certain number of 'points', say 20, redeemable at the end of the experimental session for real money. Each subject then places some fraction of his points in a 'common account,' and the remainder in the subject's 'private account'. The experimenter then tells the subjects how many points were contributed to the common account, and adds to the private account of each subject some fraction, say 40%, of the total amount in the common account. So if a subject contributes his whole 20 points to the common account, each of the four group members will receive 8 points at the end of the round. In effect, by putting the whole endowment into the common account, a player loses 12 points but the other three group members gain in total 24 (= 8 × 3) points. The players keep whatever is in their private account at the end of the round.

A self-regarding player will contribute nothing to the common account. However, only a fraction of subjects actually conform to the self-regarding model. Subjects begin by contributing on average about half of their endowment to the public account. The level of contributions decays over the course of the 10 rounds, until in the final rounds most players are behaving in a self-regarding manner (Dawes and Thaler, 1988; Ledyard, 1995). In a meta-study of 12 public goods experiments Fehr and Schmidt (1999) found that in the early rounds, average and median contribution levels ranged from 40% to 60% of the endowment, but in the final period 73% of all individuals (n = 1042) contributed nothing, and many of the remaining players contributed close to zero. These results are not compatible with the self-interested actor model, which predicts zero contribution on all rounds, though they might be predicted by a reciprocal altruism model, since the chance to reciprocate declines as the end of the experiment approaches. However, this is not in fact the explanation of moderate but deteriorating levels of cooperation in the public goods game.

The explanation of the decay of cooperation offered by subjects when debriefed after the experiment is that cooperative subjects became angry at others who contributed less than themselves, and retaliated against free-riding low contributors in the only way available to them—by lowering their own contributions (Ostrom *et al.*, 1994; Andreoni, 1995).

Experimental evidence supports this interpretation. When subjects are allowed to punish non-contributors, they do so at a cost to themselves (Orbell *et al.*, 1986; Sato, 1987; Yamagishi, 1988a,b, 1992). For instance, in Ostrom *et al.* (1992) subjects interacted for 25 periods in a public goods game, and by paying a 'fee,' subjects could impose costs on other subjects by 'fining' them. Since fining costs the individual who uses it, but the benefits of increased compliance accrue to the group as a whole, we might expect a self-regarding player to refrain from punishing. However, even a self-regarding player might engage in *strategic punishment*, expecting that by punishing in early rounds, the level of cooperation would rise enough in later rounds to render the punishing profitable. The experimenters found a significant level of punishing behaviour, but their experimental protocols made it impossible to say whether this was due to strategic punishment or strong reciprocity, since subjects were not told in advance how many periods of play they would undergo, precisely to avoid 'endgame effects'.

This shortcoming was addressed by Fehr and Gächter (2000), who set up an experimental situation in which *the possibility of strategic punishment was removed.* They used six and ten round public goods games with groups of size four, and with costly punishment allowed at the end of each round, employing three different methods of assigning members to groups. There were sufficient subjects to run between 10 and 18 groups simultaneously. Under the 'Partner' treatment, the four subjects remained in the same group for all 10 periods. Under the 'Stranger' treatment, the subjects were randomly reassigned after each round. Finally, under the 'Perfect Stranger' treatment the subjects were randomly

reassigned and assured that they would never meet the same subject more than once. Subjects earned an average of about $35 for an experimental session.

Fehr and Gächter (2000) performed their experiment for 10 rounds with punishment and 10 rounds without [for additional experimental results and their analysis, see Bowles and Gintis (2002) and Fehr and Gächter (2002).] Their results are illustrated in Figure 42.2. We see that when costly punishment is permitted, cooperation does not deteriorate, and in the Partner game, despite strict anonymity, cooperation increases almost to full cooperation, even on the final round. When punishment is not permitted, however, the same subjects experience the deterioration of cooperation found in previous public goods games. The contrast in cooperation rates between the Partner and the two Stranger treatments is worth noting, because the strength of punishment is roughly the same across all treatments. This suggests that the credibility of the punishment threat is greater in the Partner treatment because in this treatment the punished subjects are certain that, once they have been punished in previous rounds, the punishing subjects are in their group. The prosociality impact of strong reciprocity on cooperation is thus more strongly manifested, the more coherent and permanent the group in question.

42.5. The evolutionary stability of strong reciprocity

Gintis (2000b) developed an analytical model showing that under plausible conditions strong reciprocity can emerge from reciprocal altruism, through group selection. The paper models cooperation as a repeated n-person public goods game (see Section 42.4) in which, under normal conditions, if agents are sufficiently forward-looking, cooperation can be sustained by the threat of ostracism (Fudenberg and Maskin, 1986; Gintis, 2000a). However, when the group is threatened with extinction or dispersal, say through war, pestilence, or famine, cooperation is most needed for survival. During such critical periods, which were common in the evolutionary history of our species, future gains from cooperation become very uncertain, since the probability that the group will dissolve becomes high. The threat of ostracism then carries little weight, and cooperation cannot be maintained if agents are self-regarding. Thus, *precisely when a group is most in need of prosocial behaviour,*

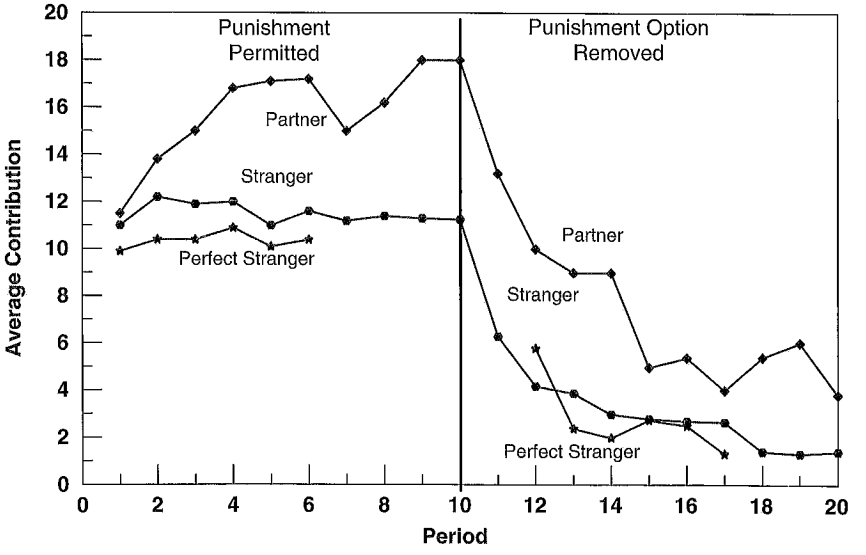

Fig. 42.2 Average contributions over time in the Partner, Stranger, and Perfect Stranger treatments when the punishment condition is played first. Adapted from Fehr and Gächter (2000).

cooperation based on reciprocal altruism will collapse.

But a small number of strong reciprocators, who punish defectors *whether or not it is in their long-term interest*, can dramatically improve the survival chances of human groups. Moreover, among species that live in groups and recognize individuals, humans are unique in their capacity to formulate and communicate rules of behaviour and to inflict heavy punishment at low cost to the punisher (Bingham, 1999), as a result of their superior tool-making and hunting ability (Goodall, 1964; Darlington, 1975; Plooij, 1978; Fifer, 1987; Isaac, 1987). Under these conditions strong reciprocators can invade a population of self-regarding types. This is because even if strong reciprocators form a small fraction of the population, at least occasionally they will form a sufficient fraction of a group such that cooperation can be maintained in bad times. Such a group will then outcompete other self-interested groups, and the fraction of strong reciprocators will grow. This will continue until an equilibrium fraction of strong reciprocators is attained.

While the above results can be obtained analytically, there is no easily interpretable mathematical expression for the equilibrium fraction of strong reciprocators. A computer simulation, however, is quite revealing. For instance, suppose in good times a group has a 95% chance of surviving one period, while in bad times (which occur one period out of 10), the group only has a 25% chance of surviving. Then the lower curve in Figure 42.3 shows the equilibrium fraction f^* of strong reciprocators as the cost of retaliation (c_r) varies and there are 40 members per group. The upper curve shows the same relationship when there are eight members per group. The latter curve would be relevant if groups are composed of a small number of 'families', and the strong reciprocity characteristic is highly transmittable within families. Note that a very small fraction of strong reciprocators can ensure cooperation, but the lower the cost of retaliation, the larger the equilibrium frequency of strong reciprocators.

This model highlights a key adaptive feature of strong reciprocity—its independence from the probability of future interactions—but it presumes that reciprocal altruism explains cooperation in normal times, when the probability of future interactions is high. However, reciprocal altruism does not work well in large groups (Taylor, 1976; Joshi, 1987; Boyd and Richerson, 1988). This is because when one withdraws cooperation in retaliation for the defection of a single group member, one inflicts punishment on all members, defector and cooperators alike. The only evolutionarily stable strategy in the n-person public goods game is to cooperate as long as all others cooperate and to defect otherwise. For any pay-off–monotonic dynamic, the basin of attraction of this equilibrium becomes very small as group size increases, so the formation of groups with a sufficient number of conditional cooperators is very unlikely, and, as a result, such an outcome may be easily disrupted by idiosyncratic play, imperfect information about the play of others, or other stochastic events. As a result, if group size is large, such an equilibrium is unlikely to be arrived at over reasonable historical time scales. Moreover, the only equilibrium is a 'knife-edge' that collapses if just one member deviates.

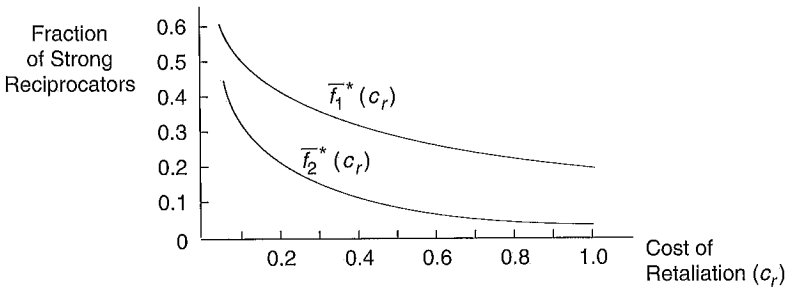

Fig. 42.3 The equilibrium fraction of strong reciprocating families: a computer simulation.

Another influential model of cooperation among self-regarding agents relies upon *reputation* effects in a repeated game setting. For instance, in *standing models* (Sugden, 1986; Boyd, 1989; Panchanathan and Boyd, 2003) individuals who are 'in good standing' in the community cooperate with others who are in good standing. If an individual fails to cooperate with someone who is in good standing, he falls into 'bad standing', and individuals in good standing will not cooperate with him. Such models are less sensitive to errors, but require that each individual know the standing of each other individual, updating with a high degree of accuracy in each period. This is plausible for small groups, but not for larger groups in which each individual observes only a small fraction of the total number of interactions among group members in each period.

In sum, strategies supporting contingent cooperation in large groups have to achieve two competing desiderata. To be stable when common, they must be intolerant of defection. But, to increase when rare there must be a substantial chance that groups with enough reciprocators can form, otherwise they cannot be evolutionarily stable, as defectors will prosper. As groups increase in size, this becomes geometrically more difficult.

To inject more realism in an evolutionary model of strong reciprocity, Henrich and Boyd (2001) developed a model in which norms for cooperation and punishment are acquired via pay-off-biased transmission (imitate the successful) and conformist transmission (imitate high-frequency behaviour). They show that if two stages of punishment are permitted, then an arbitrarily small amount of conformist transmission will stabilize cooperative behaviour by stabilizing punishment. They then explain how, once cooperation is stabilized in one group, it may spread through a multi-group population via cultural group selection. Once cooperation is prevalent, they show how prosocial genes favouring cooperation and punishment may invade in the wake of cultural group selection, for instance, because such genes decrease an individual's chance of suffering costly punishment.

This analysis reveals a deep asymmetry between altruistic cooperation and altruistic punishment, explored further in Boyd *et al.* (2003), who show that altruistic punishment allows cooperation in quite larger groups because the pay-off disadvantage of altruistic cooperators relative to defectors is independent of the frequency of defectors in the population, while the cost disadvantage of those engaged in altruistic punishment declines as defectors become rare. Thus, when altruistic punishers are common, selection pressures operating against them are weak. The fact that punishers experience only a small disadvantage when defectors are rare means that weak within-group evolutionary forces, such as conformist transmission, can stabilize punishment and allow cooperation to persist. Computer simulations show that selection among groups leads to the evolution of altruistic punishment when it could not maintain altruistic cooperation.

42.6. **Gene–culture coevolution**

If group selection is part of the explanation of the evolutionary success of cooperative individual behaviours, then it is likely that group-level characteristics, such as relatively small group size, limited migration, or frequent inter-group conflicts, that enhance group selection pressures coevolved with cooperative behaviours. Thus, group-level characteristics and individual behaviours may have synergistic effects. This being the case, cooperation is based in part on the distinctive capacities of humans to construct cultural forms that reduce phenotypic variation within groups, thus heightening the relative importance of between-group competition, and hence allowing individually costly but within-group-beneficial behaviours to coevolve with these supporting environments through a process of inter-demic group selection. The idea that the suppression of within-group competition may be a strong influence on evolutionary dynamics has been widely recognized in eusocial insects and other species. Boehm (1982) and Eibl-Eibesfeldt (1982) first applied this reasoning to human evolution, exploring the role of culturally transmitted practices that reduce phenotypic variation within groups. Examples of such practices are levelling institutions, such as monogamy and food sharing among non-kin, namely those which reduce within-group differences in reproductive fitness or material well-being. By reducing within-group differences in individual success, such structures may have attenuated within-group

genetic or cultural selection operating against individually costly but group-beneficial practices, thus giving the groups adopting them advantages in inter-group contests. Group-level institutions thus are constructed environments capable of imparting distinctive direction and pace to the process of biological evolution and cultural change. Hence, the evolutionary success of social institutions that reduce phenotypic variation within groups may be explained by the fact that they retard selection pressures working against in-group beneficial individual traits and the fact that high frequencies of bearers of these traits reduce the likelihood of group extinctions. We have modelled an evolutionary dynamic along these lines, exploring the possibility that inter-group contests play a decisive role in group-level selection. Our models assume that genetically and culturally transmitted individual behaviours, as well as culturally transmitted group-level characteristics, are subject to selection (Bowles, 2001; Bowles *et al.*, 2003). We show that inter-group conflicts may explain the evolutionary success of both: (a) altruistic forms of human sociality towards non-kin; and (b) group-level institutional structures such as food sharing and monogamy which have emerged and diffused repeatedly in a wide variety of ecologies during the course of human history. In-group-beneficial behaviours may evolve if (i) they inflict sufficient costs on out-group individuals and (ii) group-level institutions limit the individual costs of these behaviours and thereby attenuate within-group selection against these behaviours.

Our simulations show that if group-level institutions implementing resource sharing or non-random pairing among group members are permitted to evolve, group-beneficial individual traits co-evolve along with these institutions, even where the latter impose significant costs on the groups adopting them. These results hold for specifications in which cooperative individual behaviours and social institutions are initially absent in the population. In the absence of these group-level institutions, however, group-beneficial traits evolve only when inter-group conflicts are very frequent, groups are small, and migration rates are low. Thus the evolutionary success of cooperative behaviours during the last few hundred thousand years of human evolution may have been a consequence of distinctive human capacities in social-institution-building (Boyd and Richerson, 2004).

42.7. Is strong reciprocity an adaptation?

Some behavioural scientists have suggested that the behaviour we have described in this chapter was individually fitness-maximizing in our hunter-gatherer past, when anonymity and one-shot interactions were, so they say, virtually non-existent (Johnson *et al.*, 2003; Trivers, 2004). The human brain, they note, is not a general-purpose information processor, but rather a set of interacting modular systems adapted to solving the particular problems faced by our species in its evolutionary history. Since the anonymous, non-repeated interactions characteristic of experimental games were not a significant part of our evolutionary history, we could not expect subjects in experimental games to behave in a fitness-maximizing manner when confronted with them. Rather, we would expect subjects to confuse the experimental environment in more evolutionarily familiar terms as a non-anonymous, repeated interaction, and to maximize fitness with respect to this reinterpreted environment. This critique, even if correct, would not lessen the importance of strong reciprocity in contemporary societies, to the extent that modern life leads individuals to face the frequent anonymous, non-repeated interactions that are characteristic of modern societies with advanced trade, communication and transportation technologies. Thus, even if strong reciprocity were not an adaptation, it could nevertheless be an important factor in explaining human cooperation today.

But we do not believe that this critique is correct. In fact, humans are well capable of distinguishing individuals with whom they are likely to have many future interactions, from others, with whom future interactions are less likely. Indeed, human subjects cooperate much more if they expect frequent future interactions than if future interactions are rare (Gächter and Falk, 2002; Keser and vanWinden, 2000). Humans with fine-tuned behavioural repertoires depending on whether they face kin or non-kin,

repeated or one-time interactors, and whether they can or cannot gain an individual reputation probably had an evolutionary advantage in our ancestral environment. The likely reason for this advantage is that humans faced many interactions where the probability of future interactions was sufficiently low to make defection worthwhile (Manson and Wrangham, 1991; Gintis, 2000b). Humans are similarly capable of recognizing when their actions are hidden from view and profiting from such situations.

42.8. Psychological and behavioural aspects of altruism: prosocial emotions and strong reciprocity

Prosocial emotions are physiological and psychological reactions that induce agents to engage in cooperative behaviours as we have defined them above. The prosocial emotions include some, such as shame, guilt, empathy, and sensitivity to social sanction, that induce agents to undertake constructive social interactions, and others, such as the desire to punish norm violators, that reduce freeriding when the prosocial emotions fail to induce sufficiently cooperative behaviour in some fraction of members of the social group (Frank, 1987; Hirshleifer, 1987).

Without the prosocial emotions, we would all be sociopaths, and human society would not exist, however strong the institutions of contract, governmental law enforcement, and reputation. Sociopaths have no mental deficit except that their capacity to experience shame, guilt, empathy, and remorse is severely attenuated or absent. They comprise 3–4% of the male population in the USA (Mealey 1995), but account for approximately 20% of the US prison population and between 33% and 80% of the population of chronic criminal offenders.

Prosocial emotions are responsible for the host of civil and caring acts that enrich our daily lives and render living, working, shopping, and travelling among strangers feasible and pleasant. Moreover, representative government, civil liberties, due process, women's rights, respect for minorities, to name a few of the key institutions without which human dignity would be impossible in the modern world, were brought about by people involved in collective action, pursuing not only their personal ends, but also a vision for all of humanity. Our freedoms and our comforts alike are based on the emotional dispositions of generations past. While we think the evidence is strong that prosocial emotions account for important forms of human cooperation, there is no universally accepted model of how emotions combine with more cognitive processes to affect behaviours. Nor is there much agreement on how best to represent the prosocial emotions that support cooperative behaviours, although Bowles and Gintis (2002) is one attempt in this direction.

42.9. The coevolution of institutions and behaviours

If group selection is part of the explanation of the evolutionary success of cooperative individual behaviours, then it is likely that group level-characteristics—such as relatively small group size, limited migration, or frequent inter-group conflicts—that enhance group selection pressures co-evolved with cooperative behaviours. Thus group-level characteristics and individual behaviours may have synergistic effects. This being the case, cooperation is based in part on the distinctive capacities of humans to construct institutional environments that limit within-group competition and reduce phenotypic variation within groups, thus heightening the relative importance of between-group competition, and hence allowing individually costly but in-group-beneficial behaviours to coevolve with these supporting environments through a process of inter-demic group selection.

The idea that the suppression of within-group competition may be a strong influence on evolutionary dynamics has been widely recognized in eusocial insects and other species. Alexander (1979), Boehm (1982) and Eibl-Eibesfeldt (1982) first applied this reasoning to human evolution, exploring the role of culturally transmitted practices that reduce phenotypic variation within groups. Examples of such practices are levelling institutions, such as resource sharing among non-kin, namely those which reduce within-group

differences in reproductive fitness or material well-being. These practices are levelling to the extent that they result in less pronounced within-group differences in material well-being or fitness than would have obtained in their absence. Thus, the fact that good hunters who are generous towards other group members may experience higher fitness than other hunters and enjoy improved nutrition (as a result of consumption smoothing) does not indicate a lack of levelling unless these practices also result in lesser fitness and worse nutrition among less successful hunters (which seems highly unlikely).

By reducing within-group differences in individual success, such practices may have attenuated within-group genetic or cultural selection operating against individually costly but group-beneficial practices, thus giving the groups adopting them advantages in inter-group contests. Group-level institutions thus are constructed environments capable of imparting distinctive direction and pace to the process of biological evolution and cultural change. Hence, the evolutionary success of social institutions that reduce phenotypic variation within groups may be explained by the fact that they retard selection pressures working against in-group-beneficial individual traits and the fact that high frequencies of bearers of these traits reduce the likelihood of group extinctions.

We have modelled an evolutionary dynamic along these lines with the novel features that genetically and culturally transmitted individual behaviours as well as culturally transmitted group-level institutional characteristics are subject to selection, with inter-group contests playing a decisive role in group-level selection (Bowles, 2001; Bowles *et al.*, 2003).

Our simulations show that if group-level institutions implementing resource sharing or non-random pairing among group members are permitted to evolve, group-beneficial individual traits coevolve along with these institutions, even where the latter impose significant costs on the groups adopting them. These results hold for specifications in which cooperative individual behaviours and social institutions are initially absent in the population. In the absence of these group-level institutions, however, group-beneficial traits evolve only when inter-group conflicts are very frequent, groups are small, and migration rates are low. Thus the evolutionary success of cooperative behaviours in the relevant environments during the first 90 000 years of anatomically modern human existence may have been a consequence of distinctive human capacities in social-institution-building.

42.10. The internalization of norms

An internal norm is a pattern of behaviour enforced in part by internal sanctions, including shame and guilt as outlined in the previous section. People follow internal norms when they value certain behaviours for their own sake, in addition to, or despite, the effects these behaviours have on personal fitness and/or perceived well-being. The ability to internalize norms is nearly universal among humans. Although widely studied in the sociology and social psychology literature (socialization theory), it has been virtually ignored outside these fields [but see Caporael *et al.* (1989) and Simon (1990)].

Socialization models have been strongly criticized for suggesting that people adopt norms independent of their perceived pay-offs. In fact, people do not always blindly follow the norms that have been inculcated in them, but at least at times treat compliance as a strategic choice (Gintis, 1975). The 'oversocialized' model of the individual presented in the sociology literature can be counteracted by adding a phenotypic copying process reflecting the fact that agents shift from lower to higher pay-off strategies (Gintis, 2003).

All successful cultures foster internal norms that enhance personal fitness, such as future-orientation, good personal hygiene, positive work habits, and control of emotions. Cultures also universally promote altruistic norms that subordinate the individual to group welfare, fostering such behaviours as bravery, honesty, fairness, willingness to cooperate, and empathy with the distress of others.

Given that most cultures promote cooperative behaviours, and if we accept the sociological notion that individuals internalize the norms that are passed to them by parents and other influential elders, it becomes easy to explain

human cooperation. If even a fraction of society internalizes the norms of cooperation and punishes freeriders and other norm violators, a high degree of cooperation can be maintained in the long run. The puzzles are two: why do we internalize norms, and why do cultures promote cooperative behaviours?

In Gintis (2003), we provide an evolutionary model in which the capacity to internalize norms develops because this capacity enhances individual fitness in a world in which social behaviour has become too complex and multifaceted to be fruitfully evaluated piecemeal through individual rational assessment. Internalization moves norms from constraints that one can treat instrumentally towards maximizing well-being, to norms that are then valued as ends rather than means. It is not difficult to show that if an internal norm is fitness enhancing, then for plausible patterns of socialization, the allele for internalization of norms is evolutionarily stable. We may then use this framework to model Herbert Simon's (1990) explanation of altruism. Simon suggested that altruistic norms could 'hitchhike' on the general tendency of internal norms to be fitness-enhancing. However, Simon provided no formal model of this process and his ideas have been widely ignored. This paper shows that Simon's insight can be analytically modelled and is valid under plausible conditions. A straightforward gene–culture coevolution argument then explains why fitness-reducing internal norms are likely to be prosocial as opposed to socially harmful: groups with prosocial internal norms will outcompete groups with antisocial, or socially neutral, internal norms.

42.11. Conclusion

Contemporary behavioural theory is the legacy of several major contributions (Hamilton, 1964; Williams, 1966; Trivers, 1971; Wilson, 1975; Maynard Smith, 1982; Dawkins, 1989; Tooby and Cosmides, 1992), all of which assumed that the relations between non-kin could be modelled using self-regarding actors. It is not surprising, then, that the most successful research in behavioural theory has been in the area of the family, kinship, and sexual relations, while the attempts to deal with the more complex interactions characteristic of social group behaviour have been less persuasive. To address this situation, we believe that more attention should be paid to: (i) the origin and nature of social emotions (including guilt, shame, empathy, ethnic identity, and ethnic hatred); (ii) the coevolution of genes and culture in human social history; (iii) the role of group structure and group conflict in human evolution; and (iv) integrating sociobiological insights into mainstream social sciences.

Acknowledgments

We would like to thank Martin Daly, Steve Frank and Margo Wilson for helpful comments, and the Santa Fe Institute and John D. and Catherine T. MacArthur Foundation for financial support.

References

Akerlof, G. A. (1982) Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, 97: 543–569.

Alexander, R. D. (1979) *Biology and Human Affairs*. University of Washington Press, Seattle.

Andreoni, J. (1995) Cooperation in public goods experiments: kindness or confusion. *American Economic Review*, 85: 891–904.

Bingham, P. M. (1999) Human uniqueness: a general theory. *Quarterly Review of Biology*, 74: 133–169.

Blau, P. (1964) *Exchange and Power in Social Life*. Wiley, NewYork.

Boehm, C. (1982) The evolutionary development of morality as an effect of dominance behavior and conflict interference. *Journal of Social and Biological Structures*, 5: 413–421.

Bowles, S. (2001) Individual interactions, group conflicts, and the evolution of preferences. In S. N. Durlauf and H. P. Young (eds) *Social Dynamics*, pp. 155–190. MIT Press, Cambridge, MA.

Bowles, S. and Gintis, H. (2002) Homo reciprocans. *Nature*, 415: 125–128.

Bowles, S., Choi, J. and Hopfensitz, A. (2003) The co-evolution of individual behaviors and social institutions. *Journal of Theoretical Biology*, 223: 135–147.

Boyd, R. (1989) Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *Journal of Theoretical Biology*, 136: 47–56.

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

Boyd, R. and Richerson, P. J. (2004) *The Nature of Cultures*. University of Chicago Press, Chicago.

Boyd, R., Gintis, H., Bowles, S. and Richerson, P. J. (2003) Evolution of altruistic punishment, *Proceedings of the National Academy of Sciences of the USA*, 100: 3531–3535.

Camerer, C. and Thaler, R. (1995) Ultimatums, dictators, and manners. *Journal of Economic Perspectives*, 9: 209–219.

Caporael, L., Dawes, R., Orbell, J. and van de Kragt, J. C. (1989) Selfishness examined: Cooperation in the absence of egoistic incentives. *Behavioral and Brain Science*, 12: 683–738.

Darlington, P. J. (1975) Group selection, altruism, reinforcement and throwing in human evolution. *Proceedings of the National Academy of Sciences of the USA*, 72: 3748–3752.

Dawes, R. M. and Thaler, R. (1988) Cooperation. *Journal of Economic Perspectives*, 2: 187–197.

Dawkins, R. (1976, 2nd edition 1989) *The Selfish Gene*. Oxford University Press, Oxford.

Eibl-Eibesfeldt, I. (1982) Warfare, man's indoctrinability and group selection. *Journal of Comparative Ethnology*, 60: 177–198.

Fehr, E. and Gächter, S. (2000) Cooperation and punishment. *American Economic Review*, 90: 980–994.

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

Fehr, E. and Schmidt, K. M. (1999) A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114: 817–868.

Fehr, E., Kirchsteiger, G. and Riedl, A. (1993) Does fairness prevent market clearing? *Quarterly Journal of Economics*, 108: 437–459.

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

Fehr, E., Kirchsteiger, G. and Riedl, A. (1998) Gift exchange and reciprocity in competitive experimental markets. *European Economic Review*, 42: 1–34.

Feldman, M. W., Cavalli-Sforza, L. L. and Peck, J. R. (1985) Gene–culture coevolution: models for the evolution of altruism with cultural transmission. *Proceedings of the National Academy of Sciences of the USA*, 82: 5814–5818.

Fifer, F. C. (1987) The adoption of bipedalism by the hominids: a new hypothesis. *Human Evolution*, 2: 135–147.

Frank, R. H. (1987) If Homo Economicus could choose his own utility function, would he want one with a conscience? *American Economic Review*, 77: 593–604.

Fudenberg, D. and Maskin, E. (1986) The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54: 533–554.

Gächter, S. and Falk, A. (2002) Reputation and reciprocity: consequences for the labour relation. *Scandinavian Journal of Economics*, 104: 1–26.

Gächter, S. and Fehr, E. (1999) Collective action as a social exchange. *Journal of Economic Behavior and Organization*, 39: 341–369.

Ghiselin, M. T. (1974) *The Economy of Nature and the Evolution of Sex*. University of California Press, Berkeley.

Gintis, H. (1975) Welfare economics and individual development: a reply to Talcott Parsons. *Quarterly Journal of Economics*, 89: 291–302.

Gintis, H. (2000a) *Game Theory Evolving*. Princeton University Press, Princeton, NJ.

Gintis, H. (2000b) Strong reciprocity and human sociality. *Journal of Theoretical Biology*, 206: 169–179.

Gintis, H. (2003) The hitchhiker's guide to altruism: genes, culture, and the internalization of norms. *Journal of Theoretical Biology*, 220: 407–418.

Gintis, H., Bowles, S., Boyd, R. and Fehr, E. (2005) *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life*. MIT Press, Cambridge, MA.

Goodall, J. (1964) Tool-using and aimed throwing in a community of free-living chimpanzees. *Nature*, 201: 1264–1266.

Güth, W. and Tietz, R. (1990) Ultimatum bargaining behavior: a survey and comparison of experimental results. *Journal of Economic Psychology*, 11: 417–449.

Hamilton, W. D. (1964) The genetical evolution of social behavior i & ii. *Journal of Theoretical Biology*, 7: 1–16,17–52.

Henrich, J. and Boyd, R. (2001) Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, 208: 79–89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. and McElreath, R. (2001) Cooperation, reciprocity and punishment in fifteen small-scale societies. *American Economic Review*, 91: 73–78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E. and Gintis, H. (2005) Economic man' in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*.

Hirshleifer, J. (1987) Economics from a biological viewpoint. In Barney, J. B. and Ouchi, W. G. (eds) *Organizational Economics*, pp. 319–371. Jossey-Bass, San Francisco.

Homans, G. (1961) *Social Behavior: Its Elementary Forms*. Harcourt Brace, New York.

Isaac, B. (1987) Throwing and human evolution. *African Archeological Review*, 5: 3–17.

Johnson, D. P., Stopka, P. and Knights, S. (2003) The puzzle of human cooperation. *Nature*, 421: 911–912.

Joshi, N. V. (1987) Evolution of cooperation by reciprocation within structured demes. *Journal of Genetics*, 66: 69–84.

Keser, C. and vanWinden, F. (2000) Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, 102: 23–39.

Ledyard, J. O. (1995) Public goods: a survey of experimental research. In Kagel, J. H. and Roth, A. E. (eds) *The Handbook of Experimental Economics*, pp. 111–194. Princeton University Press, Princeton, NJ.

Manson, J. H. and Wrangham, R.W. (1991) Intergroup aggression in chimpanzees. *Current Anthropology*, 32: 369–390.

Maynard Smith, J. (1976) Sexual selection and the handicap principle. *Journal of Theoretical Biology*, 57: 239–242.

Maynard Smith, J. (1982) *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK.

Mealey, L. (1995) The sociobiology of sociopathy. *Behavioral and Brain Sciences*, 18: 523–541.

Orbell, J. M., Dawes, R. M. and Van de Kragt, J. C. (1986) Organizing groups for collective action. *American Political Science Review*, 80: 1171–1185.

Ostrom, E., Gardner, R. and Walker, J. (1994) *Rules, Games, and Common-Pool Resources*. University of Michigan Press, Ann Arbor.

Ostrom, E., Walker, J. and Gardner, R. (1992) Covenants with and without a sword: self governance is possible. *American Political Science Review*, 86: 404–417.

Panchanathan, K. and Boyd, R. (2003) A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology*, 224: 115–126.

Plooij, F. X. (1978) Tool-using during chimpanzees' bushpig hunt. *Carnivore*, 1: 103–106.

Rogers, A. R. (1990) Group selection by selective emigration: the effects of migration and kin structure. *American Naturalist*, 135: 398–413.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M. and Zamir, S. (1991) Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *American Economic Review*, 81: 1068–1095.

Sato, K. (1987) Distribution and the cost of maintaining common property resources. *Journal of Experimental Social Psychology*, 23: 19–31.

Simon, H. (1990) A mechanism for social selection and successful altruism. *Science*, 250: 1665–1668.

Sober, E. and Wilson, D. S. (1998) *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, Cambridge, MA.

Sugden, R. (1986) *The Economics of Rights, Co-operation and Welfare*. Blackwell, Oxford.

Taylor, M. (1976) *Anarchy and Cooperation*. Wiley, London.

Tooby, J. and Cosmides, L. (1992) The psychological foundations of culture. In J. H. Barkow, L. Cosmides and J. Tooby (eds) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 19–136. Oxford University Press, NewYork.

Trivers, R. L. (1971) Mutual benefits at all levels of life. *Science*, 304: 964–965.

Trivers, R. L. (2004) The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46: 35–57.

Williams, G. C. (1966) *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton University Press, Princeton, NJ.

Wilson, E. O. (1975) *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA.

Yamagishi, T. (1986) The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51: 110–116.

Yamagishi, T. (1988a) The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly*, 51: 265–271.

Yamagishi, T. (1988b) Seriousness of social dilemmas and the provision of a sanctioning system. *Social Psychology Quarterly*, 51: 32–42.

Yamagishi, T. (1992) Group size and the provision of a sanctioning system in a social dilemma. In W. Liebrand, D. M. Messick and H. Wilke (eds) *Social Dilemmas: Theoretical Issues and Research Findings*, pp. 267–287. Pergamon Press, Oxford.