The nature of human altruism

Ernst Fehr & Urs Fischbacher

University of Zürich, Institute for Empirical Research in Economics, Blümlisalpstrasse 10, CH-8006 Zürich, Switzerland

Some of the most fundamental questions concerning our evolutionary origins, our social relations, and the organization of society are centred around issues of altruism and selfishness. Experimental evidence indicates that human altruism is a powerful force and is unique in the animal world. However, there is much individual heterogeneity and the interaction between altruists and selfish individuals is vital to human cooperation. Depending on the environment, a minority of altruists can force a majority of selfish individuals to cooperate or, conversely, a few egoists can induce a large number of altruists to defect. Current gene-based evolutionary theories cannot explain important patterns of human altruism, pointing towards the importance of both theories of cultural evolution as well as gene–culture co-evolution.

Human societies represent a huge anomaly in the animal world[1]. They are based on a detailed division of labour and cooperation between genetically unrelated individuals in large groups. This is obviously true for modern societies with their large organizations and nation states, but it also holds for hunter-gatherers, who typically have dense networks of exchange relations and practise sophisticated forms of food-sharing, cooperative hunting, and collective warfare[2,3]. In contrast, most animal species exhibit little division of labour and cooperation is limited to small groups. Even in other primate societies, cooperation is orders of magnitude less developed than it is among humans, despite our close, common ancestry. Exceptions are social insects such as ants and bees, or the naked mole rat; however, their cooperation is based on a substantial amount of genetic relatedness.

Why are humans so unusual among animals in this respect? We propose that quantitatively, and probably even qualitatively, unique patterns of human altruism provide the answer to this question. Human altruism goes far beyond that which has been observed in the animal world. Among animals, fitness-reducing acts that confer fitness benefits on other individuals are largely restricted to kin groups; despite several decades of research, evidence for reciprocal altruism in pair-wise repeated encounters[4,5] remains scarce[6–8]. Likewise, there is little evidence so far that individual reputation building affects cooperation in animals, which contrasts strongly with what we find in humans. If we randomly pick two human strangers from a modern society and give them the chance to engage in repeated anonymous exchanges in a laboratory experiment, there is a high probability that reciprocally altruistic behaviour will emerge spontaneously[9,10].

However, human altruism extends far beyond reciprocal altruism and reputation-based cooperation, taking the form of strong reciprocity[11,12]. Strong reciprocity is a combination of altruistic rewarding, which is a predisposition to reward others for cooperative, norm-abiding behaviours, and altruistic punishment, which is a propensity to impose sanctions on others for norm violations. Strong reciprocators bear the cost of rewarding or punishing even if they gain no individual economic benefit whatsoever from their acts. In contrast, reciprocal altruists, as they have been defined in the biological literature[4,5], reward and punish only if this is in their long-term self-interest. Strong reciprocity thus constitutes a powerful incentive for cooperation even in non-repeated interactions and when reputation gains are absent, because strong reciprocators will reward those who cooperate and punish those who defect.

The first part of this review is devoted to the experimental evidence documenting the relative importance of repeated encounters, reputation formation, and strong reciprocity in human altruism. Throughout the paper we rely on a behavioural—in contrast to a psychological[13]—definition of altruism as being costly acts that confer economic benefits on other individuals. The role of kinship in human altruism is not discussed because it is well-known that humans share kin-driven altruism with many other animals[14,15]. We will show that the interaction between selfish and strongly reciprocal individuals is essential for understanding of human cooperation. We identify conditions under which selfish individuals trigger the breakdown of cooperation, and conditions under which strongly reciprocal individuals have the power to ensure widespread cooperation. Next we discuss the limits of human altruism that arise from the costs of altruistic acts. Finally, we discuss the evolutionary origins of the different forms of human altruism. We are particularly interested in whether current evolutionary models can explain why humans, but not other animals, exhibit large-scale cooperation among genetically unrelated individuals, and to what extent the evidence supports the key aspects of these models.

Proximate patterns

Altruistic behaviour in real-life circumstances can almost always be attributed to different motives. Therefore, sound knowledge about the specific motives behind altruistic acts predominantly stems from laboratory experiments. In the following, we first discuss experiments in which interactions among kin, repeated encounters, and reputation formation have been ruled out. Next, we document how the possibility of future encounters and individual reputation formation changes subjects' behaviour. In all experiments discussed below, real money, sometimes up to three months' income[16–18], was at stake. Subjects never knew the personal identities of those with whom they interacted and they had full knowledge about the structure of the experiment—the available sequence of actions and the prevailing information conditions. If, for example, the experiment ruled out future encounters between the same individuals, subjects were fully informed about this. To rule out any kind of social pressure, the design of the experiment even ensured in several instances that the experimenter could not observe subjects' individual actions but only the statistical distribution of actions[19,20].

Altruistic punishment

The ultimatum game[21] nicely illustrates that a sizeable number of people from a wide variety of cultures[22,23] even when facing high monetary stakes[16,17], are willing to punish others at a cost to themselves to prevent unfair outcomes or to sanction unfair behaviour. In this game, two subjects have to agree on the division of a fixed sum of money. Person A, the proposer, can make exactly one proposal of how to divide the money. Then person B, the responder, can accept or reject the proposed division. In the case of rejection, both receive nothing, whereas in the case of acceptance, the proposal is implemented. A robust result in this experiment is

© 2003 Nature Publishing Group

that proposals giving the responder shares below 25% of the available money are rejected with a very high probability. This shows that responders do not behave to maximize self-interest, because a selfish responder would accept any positive share. In general, the motive indicated for the rejection of positive, yet 'low', offers is that responders view them as unfair. Most proposers seem to understand that low offers will be rejected. Therefore, the equal split is often the modal offer in the ultimatum game. The decisive role of rejections is indicated by the dictator game, in which the proposer unilaterally dictates the division of the money because the responder cannot reject the offer. The average amount given to the responders in the dictator game is much lower than that in the ultimatum game[20,24].

Rejections in the ultimatum game can be viewed as altruistic acts because most people view the equal split as the fair outcome. Thus, a rejection of a low offer is costly for the responder and it punishes the proposer for the violation of a social norm. As a consequence, the proposer is likely to obey the norm in the future by making less greedy offers. For the purpose of this review, we ran an experiment with ten proposers who met a different responder in ten successive rounds. We observed that proposers who experienced a rejection in the previous round increased their offers in the current round by 7% of the available sum of money.

In the ultimatum game, the proposer's action directly affects the responder. However, a key element of the enforcement of many social norms, such as food-sharing norms in hunter-gatherer societies[2,3], is that people punish norm violators not for what they did to the punisher but for what they did to others[25,26]. Norm enforcement involves the punishment of norm violations even by those who are not economically affected by the violation. To study this question experimentally, we conducted a third-party punishment game involving three subjects—an allocator, a recipient, and a third party.[27] The allocator is endowed with 100 monetary units (MUs), the recipient has no endowment, and the third party is endowed with 50 MUs. The allocator is free to give whatever he wants to the 'poor' recipient. After the third party has been informed about the allocator's transfer to the recipient, he can spend money to punish the allocator. Every MU spent on punishment reduces the allocator's income by three MUs. Because it is costly to punish, no selfish third party will ever punish. But if a fairness norm applies to the situation, altruistic punishers are expected to punish unfair transfers. In fact, 55% of the third parties punish the allocator for transfers below 50, and the lower the transfer, the higher the punishment (Fig. 1). Moreover, between 70 and 80% of the recipients expect that allocators will be punished for unfairly low transfers. Similar results have been observed when third parties are given the chance to punish subjects in a 'prisoners' dilemma'[27]. In this case, they frequently punish a defector if his opponent cooperated. If it is anticipated, the punishment by third parties thus deters non-cooperation.

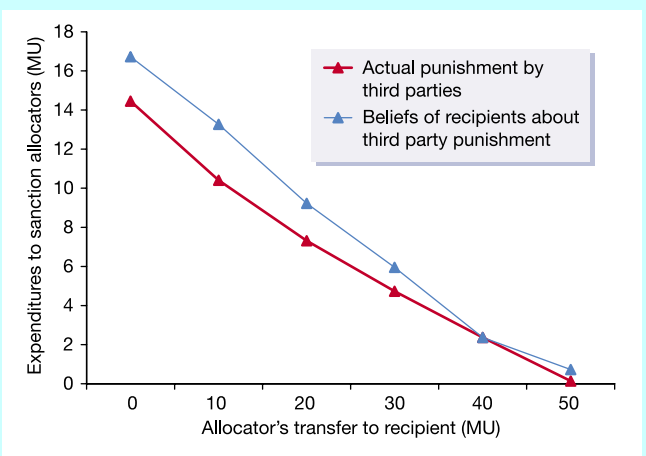

Figure 1 Altruistic punishment by third parties who are not directly affected by the violation of a fairness norm (based on ref. 27). The fair transfer level is given by 50 MUs. The more the allocator's transfer falls short of the fair level of 50 MUs, the more third parties punish the allocator. The recipients of the transfer also expect that the allocators will be punished for unfair transfers.

Altruistic rewarding

Sequentially played social dilemmas are a powerful tool for the study of altruistic rewarding. They come in various forms—as gift exchange games[28], trust games[29] or sequentially played prisoners' dilemmas[30]—but the basic structure is captured by the following example. There is a truster and a trustee, both of whom are endowed with 10 MUs. First, the truster decides how many, if any, MUs to transfer to the trustee. Then the trustee decides how much of his endowment to send to the truster. The experimenter doubles any amount sent to the other subject so that, collectively, the two subjects are best off if both transfer their whole endowment: if both keep what they have, each one earns 10; if both transfer their whole endowment, each earns 20. However, a selfish trustee will transfer nothing regardless of how much he received and, therefore, a selfish truster who anticipates this behaviour will never transfer anything in the first place.

This experiment mimics the essence of a vast number of real-life situations. A similar structure characterizes any sequential exchange that takes place in the absence of contracts that are enforced by the courts. In these situations, both players are better off exchanging their goods and favours but there is also a strong temptation to cheat. Despite the incentive to cheat, however, more than 50% of the trustees transfer money and their transfers are the higher the more the truster transferred initially[28–30]. Like altruistic punishment, the presence of altruistic rewarding has also been documented in many different countries[31], in populations with varying demographic characteristics[32], and under stake levels approaching 2–3 months' income[18].

Strong reciprocity and multilateral cooperation

A decisive feature of hunter-gatherer societies is that cooperation is not restricted to bilateral interactions. Food-sharing, cooperative hunting, and warfare involve large groups of dozens or hundreds of individuals[1]. To what extent does strong reciprocity contribute to cooperation in public goods situations involving larger groups of individuals? By definition, a public good can be consumed by every group member regardless of the member's contribution to the good. Therefore, each member has an incentive to free-ride on the contributions of others. Altruistic rewarding in this situation implies that an individual's contributions increase if the expected contributions from the other group members increase. Individuals reward others if the latter are expected to raise their cooperation.

In public goods experiments that are played only once, subjects typically contribute between 40 and 60% of their endowment, although selfish individuals are predicted to contribute nothing[33]. There is also strong evidence that higher expectations about others' contributions induce individual subjects to contribute more[33–35]. Cooperation is, however, rarely stable and deteriorates to rather low levels if the game is played repeatedly (and anonymously) for ten rounds[36,37].

The most plausible interpretation of the decay of cooperation is based on the fact that a large percentage of the subjects are strong reciprocators but that there are also many total free-riders who never contribute anything[35]. Owing to the existence of strong reciprocators, the 'average' subject increases his contribution levels in response to expected increases in the average contribution of other group members. Yet, owing to the existence of selfish subjects, the intercept and the steepness of this relationship is insufficient to establish an equilibrium with high cooperation (Fig. 2). In round

© 2003 Nature Publishing Group

one, subjects typically have optimistic expectations about others' cooperation but, given the aggregate pattern of behaviours, this expectation will necessarily be disappointed, leading to a breakdown of cooperation over time.

This breakdown of cooperation provides an important lesson. Despite the fact that there are a large number of strong reciprocators, they cannot prevent the decay of cooperation under these circumstances. In fact, it can be shown theoretically that in a population with a clear majority of strong reciprocators, a small minority of selfish individuals suffices to render zero cooperation the unique equilibrium[38]. This implies that it is not possible to infer the absence of altruistic individuals from a situation in which we observe little cooperation. If strong reciprocators believe that no one else will cooperate, they will also not cooperate. To maintain cooperation in *n*-person interactions, the upholding of the belief that all or most members of the group will cooperate is thus decisive.

Any mechanism that generates such a belief has to provide cooperation incentives for the selfish individuals. The punishment of non-cooperators in repeated interactions[39–41] or altruistic punishment[27,42] provide two such possibilities. If cooperators have the opportunity to target their punishment directly towards those who defect they impose strong sanctions on the defectors. Thus, in the presence of targeted punishment opportunities, strong reciprocators are capable of enforcing widespread cooperation by deterring potential non-cooperators[39,40,42]. In fact, it can be shown theoretically that even a minority of strong reciprocators suffices to discipline a majority of selfish individuals when direct punishment is possible[38].

Repeated interactions and reputation formation

A reputation for behaving altruistically is another powerful mechanism for the enforcement of cooperation in public goods situations. If people are engaged in bilateral encounters as well as in *n*-person public goods interactions, a defection in the public goods situation, if known by others, may decrease others' willingness to help in bilateral interactions[43]. Suppose that after each round of interaction

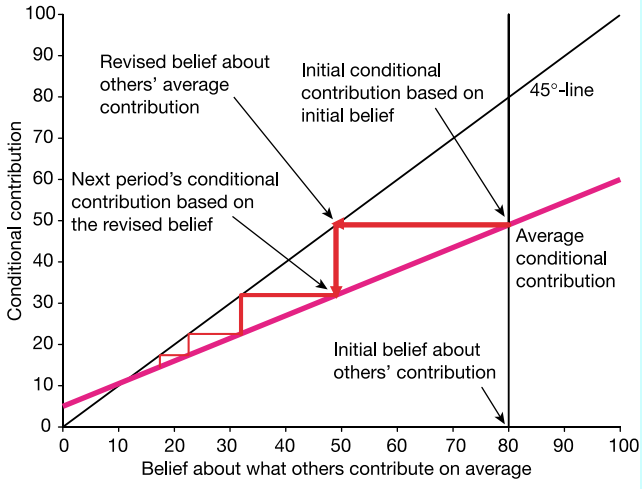

Figure 2 The decay of cooperation over time. Subjects are heterogeneous with regard to their willingness to reward altruistically. This results in the relationship between the expected average contribution of other group members to the public good and the contribution of a representative individual (the average conditional contribution indicated by the purple line). Initially, individuals expect high average contribution rates, say 80% of the endowment. On average, this induces them to contribute 50%. Therefore, expectations are disappointed which leads to a downwards revision of expectations to say, 50% of the endowment. Yet, if individuals expect 50% they will in fact only contribute roughly 30%, causing a further downwards revision of expectations. The process stops at the intersection point with the 45° line, which determines the equilibrium level of altruistic cooperation in this setting.

in a public goods experiment, subjects play an indirect reciprocity game[44]. In this game, subjects are matched in pairs and one subject is randomly placed in the role of a donor and the other in the role of a recipient. The donor can help the recipient, and the donor's costs of helping are lower than the benefits for the recipient. The recipient's reputation is established by his decision in the previous public goods round and his history of helping decisions in the indirect reciprocity game. It turns out that the recipients' reputations in the public goods game are an important determinant for the donors' decisions. Donors punish the recipients by being significantly less likely to help when the recipients defected in the previous public goods game. This, in turn, has powerful cooperation-enhancing effects in the future rounds of the public goods game.

Helping behaviour in indirect reciprocity experiments has also been documented in the absence of interactions in public goods games[45,46]. A crucial element in these experiments is that direct reciprocity is ruled out because no recipient will ever be put in a position where he can give to one of his previous donors. Helping rates between 50 and 90% have been achieved and recipients with a history of generous helping decisions are significantly more likely to receive help themselves. This suggests that the donors' behaviour may be driven by the desire to acquire a good reputation. However, it is also possible that altruistic rewarding drives helping behaviour. A recent study examines this question by allowing only half of the subjects in the experiment to acquire a reputation[47]. This means that one can compare the behaviour of donors who cannot gain a reputation with the behaviour of those who can. The data show that both altruistic rewarding and reputation-seeking are powerful determinants of donors' behaviour. Donors who cannot acquire a reputation help in 37% of the cases whereas those who can gain a reputation help in 74% of the cases. These results indicate that humans are very attentive to possibilities of individual reputation formation in the domain of rewarding behaviours. They exhibit a sizeable baseline level of altruistic rewarding, and when given the opportunity to gain a reputation for being generous, helping rates increase strongly. Humans are similarly attentive to the possibility of repeated interactions with the same individual (reciprocal altruism). The cooperation rate is much higher in social dilemmas if subjects know that there is a possibility of meeting the same partners again in future periods[10].

Little is known about repetition and reputation effects in the domain of punishing behaviours. We conducted a series of ten ultimatum games in two conditions for this purpose—a reputation condition and a baseline condition. In both conditions, 10 MUs have to be divided in every period and every proposer is matched with a new responder in each of the ten games. In the reputation condition, the proposers are informed about the current responder's past rejection behaviour, whereas this knowledge is absent in the baseline condition. This means that the responders in the reputation condition can gain an individual reputation for being tough bargainers by rejecting even high offers. A responder who incurs the short-term cost of a rejection can gain the long-term benefits of a 'good' reputation by inducing future proposers to make him better offers. Because this economic benefit is absent in the baseline condition, subjects who understand the logic of reputation formation will exhibit higher acceptance standards in the reputation condition.

In both conditions, the responders indicated an acceptance threshold in each period, that is, the smallest amount they were willing to accept. The results show that when the subjects were in the baseline condition first, the average acceptance threshold was about 3 MUs, whereas if they entered the reputation condition, their thresholds immediately jumped to more than 4 MUs (Fig. 3a). This jump in the thresholds forced the proposers to increase their offers. Similarly, if the reputation condition took place first, the average thresholds immediately decreased when subjects entered the baseline condition (Fig. 3a). Moreover, this change in the average

© 2003 Nature Publishing Group

thresholds is not an artefact of aggregation. It is explained by the fact that the vast majority of responders (82%, $n = 94$) increase the threshold in the reputation condition relative to the baseline (Fig. 3b) while the remaining minority keep the thresholds roughly constant. These results suggest that altruistic punishers clearly understand that if individual reputation building is possible, it pays to acquire a reputation as a tough bargainer. This also means that their rejections in the baseline condition cannot be attributed to cognitive problems in understanding when individual reputation matters and when it does not.

Limits of human altruism

Strongly reciprocal individuals reward and punish in anonymous one-shot interactions. Yet they increase their rewards and punishment in repeated interactions or when their reputation is at stake. This suggests that a combination of altruistic and selfish concerns motivates them. Their altruistic motives induce them to cooperate and punish in one-shot interactions and their selfish motives induce them to increase rewards and punishment in repeated interactions or when reputation-building is possible. If this argument is correct, we should also observe that altruistic acts become less frequent as their costs increase. At a higher cost, individuals have to give up more of their own payoff to help others, so that the individuals will exhibit less altruistic behaviour for a given combination of selfish and altruistic motives. The evidence from dictator games and public good games confirms this prediction. If the own payoff that needs to be invested to produce one unit of the public good increases, subjects invest less into the public good[36,37]. Likewise, if the cost of transferring one MU to the recipient in the dictator game increases, the dictators give less money to the recipients[48].

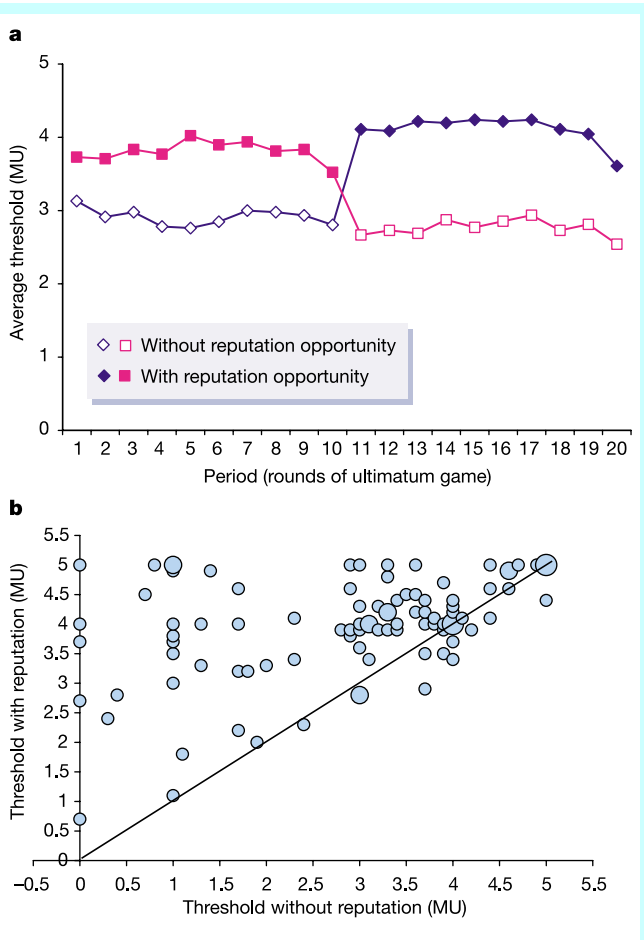

Figure 3 Responders' acceptance thresholds in the ultimatum game with and without reputation opportunities. **a**, Time trend of acceptance thresholds. If the control treatment without the opportunity to build an individual reputation for toughness is conducted first, the responders reject offers below 3 MUs (open blue symbols). Immediately after the implementation of reputation building opportunities in period 11, the acceptance thresholds jump up to more than 4 MUs, indicating the desire to be known as a 'tough' responder (solid blue symbols). If the reputation treatment comes first (purple symbols) the removal of the opportunity to acquire a reputation immediately causes a decrease in responders' acceptance thresholds. **b**, Individual level changes in responders' average acceptance thresholds. The relative size of the circles represents the frequency of observations behind a circle. Responders who increase their average acceptance threshold in the reputation condition relative to the baseline condition generate a data point above the 45° line. The vast majority of responders increase their thresholds when they can gain a reputation for toughness. Only a small minority lowers the thresholds or keeps them roughly constant.

Proximate theories

Altruistic rewards and punishment imply that individuals have proximate motives beyond their economic self-interest—their subjective evaluations of economic payoffs differ from the economic payoffs. Although this is an old idea[49], formal theories of non-selfish motives with predictive power in a wide range of circumstances have only recently been developed. These theories formalize notions of inequity aversion[38,50] and reciprocal fairness[51–53]. They predict, for example, that many subjects in the prisoners' dilemma prefer mutual cooperation over unilateral defection, even though it is in their economic self-interest to defect regardless of what the other player does. This prediction is supported by the evidence[30,54] and has wide-ranging implications. If the players have such preferences, the game is no longer a prisoners' dilemma but an assurance game in which both mutual defection as well as mutual cooperation are equilibria. The crucial point is that such subjects are willing to cooperate if they believe that their opponent will cooperate and, therefore, mutual cooperation is an equilibrium. However, because mutual defection is also an equilibrium, it depends on the individuals' beliefs about the other players' actions as to whether the mutual cooperation or the mutual defection equilibrium is played.

Recent results on the neurobiology of cooperation in the prisoners' dilemma support the view that individuals experience particular subjective rewards from mutual cooperation[55]. If subjects achieve the mutual cooperation outcome with another human subject, the brain's reward circuit (components of the mesolimbic dopamine system including the striatum and the orbitofrontal cortex) is activated relative to a situation in which subjects achieve mutual cooperation with a programmed computer. Moreover, there is also evidence indicating a negative response of the dopamine system if a subject cooperates but the opponent defects.

Evolutionary origins

What are the ultimate origins behind the rich patterns of human altruism described above? It must be emphasized in the context of this question that a convincing explanation of the distinct features of human altruism should be based on capacities which are distinctly human—otherwise there is the risk of merely explaining animal, not human, altruism.

Reciprocal altruism

Reciprocal altruism[4,5] in the form of tit-for-tat or similar cooperation-sustaining strategies in the repeated prisoners' dilemma is a powerful ultimate explanation for human altruism in small and stable groups. The experimental evidence unambiguously shows that subjects cooperate more in two-person interactions if future interactions are more likely[9,10]. There are, however, several aspects of human interactions that point towards the need to go beyond reciprocal altruism: first, with a few exceptions[26,56], the evolutionary analysis of repeated encounters has been largely restricted to two-person interactions but the human case clearly demands the analysis of larger groups. Unfortunately, the

© 2003 Nature Publishing Group

evolutionary success of tit-for-tat-like strategies of conditional cooperation is extremely limited even in relatively small groups of 4–8 individuals. It turns out that in a repeated *n*-person prisoners' dilemma, the only conditionally cooperative, evolutionarily stable strategy prescribes cooperation only if all other group members cooperated in the previous period. The basin of attraction of this strategy is extremely small because a few selfish players suffice to undermine the cooperation of the conditional cooperators.[56]

Second, the interacting individuals are forced to stay together for a random number of periods[6]. This assumption is not only violated by many, if not most, animal interactions but it is also clearly violated in the case of humans. Throughout evolutionary history, humans almost always had the option to stop interacting with genetically unrelated individuals. Thus, the choice of cooperation partners has to be modelled explicitly, which brings issues of reputation formation to the forefront of the analysis. Recent experiments indicate that endogenous partner choice and the associated incentives for reputation formation have powerful effects on human cooperation[57].

Third, reciprocal altruism relies on the idea that altruistic behaviour creates economic benefits for the altruist in the future. Therefore, it has difficulties explaining strongly reciprocal behaviour, which is characterized by the absence of future net benefits. Reciprocal altruism could only explain strong reciprocity if humans behave as if cooperation provides future net benefits, although there are none objectively. The ethnographic evidence suggests that—depending on the interaction partner—humans faced many different probabilities of repeated encounters so that situations often arose in which defection was the best strategy[58]. Indeed, the very fact that humans seem to have excellent cheating detection abilities[59] suggests that, despite many repeated interactions, cheating has been a major problem throughout human evolution. Therefore, humans' behavioural rules are likely to be fine-tuned to the variations in cheating opportunities, casting doubt on the assumption that humans systematically overestimate the future benefits from current altruistic behaviours.

Reputation-seeking

Evolutionary approaches to reputation-based cooperation[44,60–64] represent important steps beyond reciprocal altruism. The indirect reciprocity model[44,60,61] relies on the idea that third parties reward individuals with an altruistic reputation if they can acquire a good reputation themselves by rewarding. It has been shown that aspects of the food-sharing pattern of the Ache of Paraguay can be explained by this logic[65]. The experimental evidence also strongly suggests that a considerable part of human altruism is driven by concerns about reputation. Yet there are still some unsolved theoretical problems that point towards the need for further research. First, the indirect reciprocity approach produces long-run helping rates of roughly 40% if the recipient's benefit is four times the donor's cost, provided that all individuals live in isolated groups without any migration. If, however, genetic mixing between the groups occurs, helping rates decline dramatically and approach zero[61]. It would be an important step forward if the indirect reciprocity approach could be modified in such a way that significant helping rates could be maintained under reasonable assumptions about migration between groups. Second, the question of how to model the concept of a good reputation remains open. For example, should an individual who does not help a person with a bad reputation lose his good reputation? Currently the image-scoring approach[44] gives an affirmative answer to this question while others do not[61].

Third, reputation formation among humans is based on our language capabilities. However, we can use our language to tell the truth or to lie. Thus, what ensures that individuals' reputations provide a reasonably accurate picture of their past behaviours? Fourth, the indirect reciprocity approach is limited to dyadic cooperation. Therefore, it cannot currently explain cooperation in larger groups. But recent experiments that connect the *n*-person public good game with an indirect reciprocity game do point towards a potential solution[43]. Finally, reputation-based approaches cannot account for strong reciprocity unless one assumes that humans behave as if they systematically overestimate the future gains from current altruistic acts—an assumption that is dubious in view of the experimental evidence.

Costly signalling theory also provides a reputation-based ultimate explanation for altruistic behaviour[62,63]. According to this approach, individuals signal favourable, yet unobservable, traits with altruistic acts, rendering them preferred mating partners or helping in the recruitment of coalition partners in conflicts. The assumption behind this theory is that individuals with better traits have lower marginal signalling costs, that is, lower costs of altruistic acts. Thus, those with better traits are more likely to signal, which allows the inference that those who signal have better traits on average. The advantage of this approach is that it could, in principle, explain contributions to *n*-person public goods. The weakness is that the signalling of unobservable traits need not occur by altruistic acts but can also take other forms. The approach, therefore, generates multiple equilibria—in some equilibria, signalling occurs via altruistic behaviour; in others, signalling does not involve any altruistic acts. Therefore, this theory has difficulties explaining human altruism unless it is supplemented with some other mechanisms. One such mechanism might be cultural group selection[63]. If groups where signalling takes place via altruistic behaviour have better survival prospects, selection between groups favours those groups which have settled at a pro-social within-group equilibrium. Since there is no within-group selection against the altruists at the pro-social equilibrium, only weak effects of cultural selection between groups are required here. There is evidence[66] from Meriam turtle hunters that is consistent with costly signalling theory but so far there is no experimental evidence for altruistic costly signalling.

Gene–culture coevolution

The birth of modern sociobiology is associated with scepticism against genetic group selection[67]; although it is possible in theory, and in spite of a few plausible cases[25], genetic group selection has generally been deemed unlikely to occur empirically. The main argument has been that it can at best be relevant in small isolated groups because migration in combination with within-group selection against altruists is a much stronger force than selection between groups. The migration of defectors to groups with a comparatively large number of altruists plus the within-group fitness advantage of defectors quickly removes the genetic differences between groups so that group selection has little effect on the overall selection of altruistic traits[68]. Consistent with this argument, genetic differences between groups in populations of mobile vertebrates such as humans are roughly what one would expect if groups were randomly mixed[69]. Thus, purely genetic group selection is, like the gene-based approaches of reciprocal altruism and indirect reciprocity, unlikely to provide a satisfactory explanation for strong reciprocity and large-scale cooperation among humans. However, the arguments against genetic group selection are far less persuasive when applied to the selection of culturally transmitted traits. Cultural transmission occurs through imitation and teaching, that is, through social learning. There are apparent large differences in cultural practices of different groups around the world and ethnographic evidence indicates that even neighbouring groups are often characterized by very different cultures and institutions[70]. In addition, a culture-based approach makes use of the human capacity to establish and transmit behavioural norms through social learning—a capacity that is quantitatively, and probably even qualitatively, distinctly human[1,71].

Recent theoretical models of cultural group selection[72,73] or of gene–culture coevolution[71,74] could provide a solution to the puzzle of strong reciprocity and large-scale human cooperation. They are

© 2003 Nature Publishing Group

based on the idea that norms and institutions—such as food-sharing norms or monogamy—are sustained by punishment and decisively weaken the within-group selection against the altruistic trait. If altruistic punishment is ruled out, cultural group selection is not capable of generating cooperation in large groups (Fig. 4). Yet, when punishment of non-cooperators and non-punishers is possible, punishment evolves and cooperation in much larger groups can be maintained[73]. This is due to the fact that the altruistic punishment of non-cooperators in combination with the imitation of economically successful behaviours prevents the erosion of group differences with regard to the relative frequency of cooperating members. If there are a sufficient number of altruistic punishers, the cooperators do better than the defectors because the latter are punished. Therefore, cooperative behaviour is more likely to be imitated. Moreover, when cooperation in a group is widespread, altruistic punishers have only a small or no within-group disadvantage relative to pure cooperators who do not punish. At the limit, when everybody cooperates, punishers incur no punishment costs at all and thus have no disadvantage. Thus, small cultural group selection effects suffice to overcome the small cost disadvantage of altruistic punishers that arises from the necessity of punishing mutant defectors.

To what extent is there evidence for the role of culture and group selection in human altruism? There is strong evidence from intergenerational ultimatum and trust games that advice from players who previously participated in the experiment increases altruistic punishment and altruistic rewarding[75]. Recent intergenerational public good games where advice is given indicate that later generations achieve significantly higher cooperation levels even in the absence of punishment opportunities[76]. Ultimatum and dictator games with children of different ages show that older children are more generous and more willing to punish altruistically[77]. Although these changes in children's behaviour could be a result of genetic developmental processes, it seems at least as plausible to assume that they are also a product of socialization by parents and peers. Why, after all, do parents invest so much time and energy into the proper socialization of their children if this effort is futile? Perhaps the strongest evidence for the role of cultural norms comes from a series of experiments in 15 small-scale societies[23], showing decisive differences across societies in the behaviour of proposers and responders in the ultimatum game. Some tribes like the Hazda from Tanzania exhibit a considerable amount of altruistic punishment whereas the Machiguenga from Peru show little concern about fair sharing. Thus, taken together, there is fairly convincing evidence that cultural forces exert a significant impact on human altruism.

Yet, what is the evidence for cultural group selection? There is quite strong evidence that group conflict and warfare were widespread in foraging societies[78,79]. There are also examples[70,80] suggesting that group conflict contributes to the cultural extinction of groups because the winning groups force their cultural norms and institutions on the losing groups. However, although these examples are suggestive, they are not conclusive, so further evidence is needed.

If cultural group selection was a significant force in evolution, then the human propensity to reward and punish altruistically should be systematically affected by inter-group conflicts. In particular, altruistic cooperation should be more prevalent if cooperative acts contribute to success in a group conflict. Likewise, people should be more willing to punish defectors if defection occurs in the context of a group conflict. There is evidence from inter-group conflict games indicating that altruistic cooperation in prisoners' dilemmas indeed increases if the game is embedded in an inter-group conflict[81]. However, there is no evidence so far showing that inter-group conflicts increase altruistic punishment.

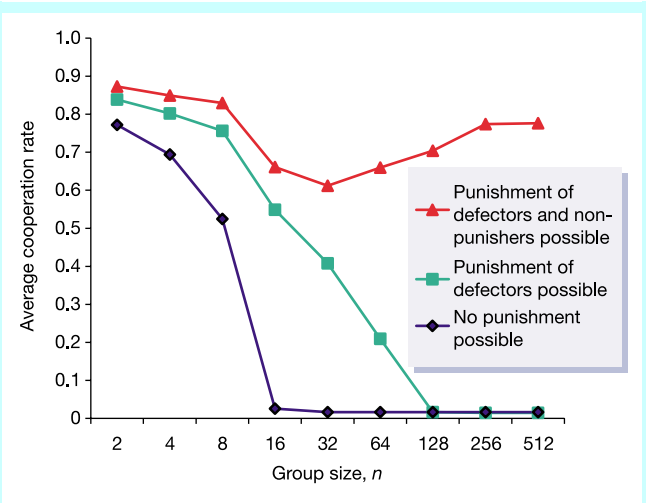

Figure 4 Simulations of the evolution of cooperation in multi-person prisoners' dilemmas with group conflicts and different degrees of altruistic punishment. The simulations are based on the model of ref. 73 but we added the possibility of punishing the non-punishers. There are 64 groups of fixed size n with n ranging from 2 to 512. The figure shows the average cooperation rate in 100 independent simulations over the last 1,000 of 2,000 generations. If the altruistic punishment of defectors is ruled out, cooperation already breaks down for groups of size 16 and larger. If altruistic punishment of defectors is possible, groups of size 32 can still maintain a cooperation rate of 40%. However, the biggest impact from altruistic punishment prevails if non-punishers can also be punished. In this case, even groups of several hundred individuals can establish cooperation rates of between 70 and 80%.

Open questions

We now know a lot more about human altruism than we did one decade ago. There is experimental evidence indicating that repeated interactions, reputation-formation, and strong reciprocity are powerful determinants of human behaviour. There are formal models that capture the subtleties of interactions between selfish and strongly reciprocal individuals, and there is a much better understanding about the nature of the evolutionary forces that probably shaped human altruism. However, there are still a considerable number of open questions. In view of the relevance of cultural evolution, it is necessary to study the relationship between cultural and economic institutions and the prevailing patterns of human altruism. Although recent evidence[23] suggests that market integration and the potential gains from cooperation are important factors, our knowledge is still extremely limited. This limitation is partly due to the fact that far too many experiments use students from developed countries as participants. Instead, we need experiments with participants that are representative of whole countries or cultures and we need to combine behavioural measures of altruism with individual-level demographic data and group-level data about cultural and economic institutions. In view of the theoretical importance of group conflicts and group reputation, much more evidence on how these affect altruistic rewarding and punishment is necessary. We also need more empirical knowledge about the characteristics of the individual reputation acquired by people and how others respond to this reputation.

At the ultimate level, the evolution and role of altruistic rewarding for cooperation in larger groups remains in the dark. Likewise, the empirical study of altruistic rewarding has been largely limited to dyadic interactions and little is known about how cooperation in n-person public good situations is affected if subjects have the opportunity to altruistically reward others after having observed each others' contribution choices. Evolutionary explanations of this kind of altruistic rewarding are likely to be much more difficult than explanations of altruistic punishment because, when cooperation is frequent, rewarding causes high costs for the altruists whereas a credible punishment threat renders actual punishment unnecessary. Finally, to enhance the study of the evolution of human altruism,

© 2003 Nature Publishing Group

there is a great need for empirically testable predictions that are rigorously derived from the evolutionary models. □

doi:10.1038/nature02043.

1. Boyd, R. & Richerson, P. *The Nature of Cultures* (Univ. Chicago Press, Chicago, in the press).
2. Kaplan, H., Hill, J., Lancaster, J. & Hurtado, A. M. A theory of human life history evolution: diet, intelligence, and longevity. *Evol. Anthropol.* **9,** 156–185 (2000).
3. Hill, K. Altruistic cooperation during foraging by the Ache, and the evolved human predisposition to cooperate. *Hum. Nat.* **13,** 105–128 (2002).
4. Trivers, R. L. Evolution of reciprocal altruism. *Q. Rev. Biol.* **46,** 35–57 (1971).
5. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211,** 1390–1396 (1981).
6. Hammerstein, P. in *Genetic and Cultural Evolution of Cooperation. Dahlem Workshop Report 90.* (ed. Hammerstein, P.) 1–11 (MIT Press, Cambridge, MA, 2003).
7. Stephens, D. W., McLinn, C. M. & Stevens, J. R. Discounting and reciprocity in an iterated prisoner's dilemma. *Science* **298,** 2216–2218 (2002).
8. Hauser, M. D., Chen, K. M., Frances, C. & Chuang, E. Give unto others: Genetically unrelated cotton-tamarin monkeys preferentially give food to those who altruistically give food back. *Proc. R. Soc. Lond. B* (in the press).
9. Andreoni, J. & Miller, J. Rational cooperation in the finitely repeated prisoner's dilemma: experimental evidence. *Econ. J.* **103,** 570–585 (1993).
10. Gächter, S. & Falk, A. Reputation and reciprocity: consequences for the labour relation. *Scand. J. Econ.* **104,** 1–26 (2002).
11. Fehr, E., Fischbacher, U. & Gächter, S. Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum. Nat.* **13,** 1–25 (2002).
12. Gintis, H. Strong reciprocity and human sociality. *J. Theor. Biol.* **206,** 169–179 (2000).
13. Batson, D. C. *The Altruism Question* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1991).
14. Silk, J. B. Adoption and kinship in Oceania. *Am. Anthropol.* **82,** 799–820 (1980).
15. Daly, M. & Wilson, M. Evolutionary social-psychology and family homicide. *Science* **242,** 519–524 (1988).
16. Cameron, L. A. Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Econ. Inq.* **37,** 47–59 (1999).
17. Slonim, R. & Roth, A. E. Learning in high stakes ultimatum games: An experiment in the Slovak republic. *Econometrica* **66,** 569–596 (1998).
18. Fehr, F., Tougareva, E. & Fischbacher, U. *Do High Stakes and Competition Undermine Fairness?* Working Paper 125 (Institute for Empirical Research in Economics, Univ. Zurich, 2002).
19. Bolton, G. & Zwick, R. Anonymity versus punishment in ultimatum bargaining. *Game Econ. Behav.* **10,** 95–121 (1995).
20. Hoffman, E., McCabe, K., Shachat, K. & Smith, V. Preferences, property rights and anonymity in bargaining games. *Game Econ. Behav.* **7,** 346–380 (1994).
21. Güth, W., Schmittberger, R. & Schwarze, B. An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Organ.* **3,** 367–388 (1982).
22. Roth, A., Prasnikar, V., Okuno-Fujiwara, M. & Zamir, S. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An experimental study. *Am. Econ. Rev.* **81,** 1068–1095 (1991).
23. Henrich, J. *et al.* In search of *Homo economicus*: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* **91,** 73–78 (2001).
24. Forsythe, R., Horowitz, J. L., Savin, N. E. & Sefton, M. Fairness in simple bargaining experiments. *Game Econ. Behav.* **6,** 347–369 (1994).
25. Sober, E. & Wilson, D. S. *Unto Others—the Evolution and Psychology of Unselfish Behavior* (Harvard Univ. Press, Cambridge, MA, 1998).
26. Bendor, J. & Swistak, P. The evolution of norms. *Am. J. Sociol.* **106,** 1493–1545 (2001).
27. Fehr, E. & Fischbacher, U. Third party punishment and social norms. *Evol. Hum. Behav.* (in the press).
28. Fehr, E., Kirchsteiger, G. & Riedl, A. Does fairness prevent market clearing? An experimental investigation. *Q. J. Econ.* **108,** 437–459 (1993).
29. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity and social history. *Game Econ. Behav.* **10,** 122–142 (1995).
30. Hayashi, N., Ostrom, E., Walker, J. & Yamagishi, T. Reciprocity, trust, and the sense of control—a cross-societal study. *Rational. Soc.* **11,** 27–46 (1999).
31. Buchan, N. R., Croson, R. T. A. & Dawes, R. M. Swift neighbors and persistent strangers: a cross-cultural investigation of trust and reciprocity in social exchange. *Am. J. Sociol.* **108,** 168–206 (2002).
32. Fehr, E., Fischbacher, U., Rosenbladt, B., Schupp, J. & Wagner, G. A nationwide laboratory—examining trust and trustworthiness by integrating behavioural experiments into representative surveys. *Schmoller Jahrbuch* **122,** 519–542 (2002).
33. Dawes, R. M. Social dilemmas. *Annu. Rev. Psychol.* **31,** 169–193 (1980).
34. Messick, D. & Brewer, M. in *Review of Personality and Social Psychology* (ed. Wheeler, L.) (Sage Publ., Beverly Hills, 1983).
35. Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* **71,** 397–404 (2001).
36. Isaac, R. M. & Walker, J. M. Group-size effects in public-goods provision—the voluntary contributions mechanism. *Q. J. Econ.* **103,** 179–199 (1988).
37. Ledyard, J. in *Handbook of Experimental Economics* (eds Kagel, J. & Roth, A.) 111–194 (Princeton Univ. Press, 1995).
38. Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114,** 817–868 (1999).
39. Yamagishi, T. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51,** 110–116 (1986).
40. Ostrom, E., Walker, J. & Gardner, R. Covenants with and without a sword: self-governance is possible. *Am. Polit. Sci. Rev.* **86,** 404–417 (1992).
41. Sethi, R. & Somanathan, E. The evolution of social norms in common property resource use. *Am. Econ. Rev.* **86,** 766–788 (1996).
42. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415,** 137–140 (2002).
43. Milinski, M., Semmann, D. & Krambeck, H. J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415,** 424–426 (2002).
44. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393,** 573–577 (1998).
45. Wedekind, C. & Milinski, M. Cooperation through image scoring in humans. *Science* **288,** 850–852 (2000).
46. Milinski, M., Semmann, D., Bakker, T. C. M. & Krambeck, H. J. Cooperation through indirect reciprocity: Image scoring or standing strategy? Proc. *R. Soc. Lond. B* **268,** 2495–2501 (2001).
47. Engelmann, D. & Fischbacher, U. *Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game* (Working Paper 132, Institute for Empirical Research in Economics, Univ. Zurich, 2002).
48. Andreoni, J. & Miller, J. Giving according to Garp: an experimental test of the consistency of preferences for altruism. *Econometrica* **70,** 737–753 (2002).
49. Thibaut, J. W. & Kelley, H. H. *The Social Psychology of Groups* (Wiley, New York, 1959).
50. Bolton, G. E. & Ockenfels, A. Erc: A theory of equity, reciprocity, and competition. *Am. Econ. Rev.* **90,** 166–193 (2000).
51. Rabin, M. Incorporating fairness into game theory and economics. *Am. Econ. Rev.* **83,** 1281–1302 (1993).
52. Levine, D. K. Modeling altruism and spitefulness in experiments. *Rev. Econ. Dynam.* **1,** 593–622 (1998).
53. Falk, A. & Fischbacher, U. Distributional consequences and intentions in a model of reciprocity. *Ann. Econ. Stat.* **63,** 111–129 (2001).
54. Kiyonari, T., Tanida, S. & Yamagishi, T. Social exchange and reciprocity: confusion or a heuristic. *Evol. Hum. Behav.* **21,** 411–427 (2000).
55. Rilling, J. K. *et al.* A neural basis for social cooperation. *Neuron* **35,** 395–405 (2002).
56. Boyd, R. & Richerson, P. J. The evolution of reciprocity in sizable groups. *J. Theor. Biol.* **132,** 337–356 (1988).
57. Brown, M., Falk, A. & Fehr, E. Relational contracts and the nature of market interactions. *Econometrica* (in the press).
58. Fehr, E. & Henrich, J. in *Genetic and Cultural Evolution of Cooperation. Dahlem Workshop Report 90* (ed. Hammerstein, P.) 55–82 (MIT Press, Cambridge, MA, 2003).
59. Cosmides, L. & Tooby, J. in *The Adapted Mind* (eds Barkow, J., Cosmides, L. & Tooby, J.) (Oxford Univ. Press, New York, 1992).
60. Alexander, R. D. *The Biology of Moral Systems* (Aldine De Gruyter, New York, 1987).
61. Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocity. *Proc. R. Soc Lond. B* **268,** 745–753 (2001).
62. Zahavi, A. Altruism as a handicap—the limitations of kin selection and reciprocity. *J. Avian Biol.* **26,** 1–3 (1995).
63. Gintis, H., Smith, E. A. & Bowles, S. Costly signaling and cooperation. *J. Theor. Biol.* **213,** 103–119 (2001).
64. Nowak, M. A., Page, K. M. & Sigmund, K. Fairness versus reason in the ultimatum game. *Science* **289,** 1773–1775 (2000).
65. Gurven, M., Allen-Arave, W., Hill, K. & Hurtado, M. It's a wonderful life: Signaling generosity among the Ache of Paraguay. *Evol. Hum. Behav.* **21,** 263–282 (2000).
66. Smith, E. A., Bliege Bird, R. L. & Bird, D. W. The benefits of costly signalling: Meriam turtle hunters. *Behav. Ecol.* **14,** 116–126 (2003).
67. Williams, G. D. *Adaption and Natural Selection: A Critique of Some Current Evolutionary Thought* (Princeton Univ. Press, Princeton, 1966).
68. Aoki, M. A condition for group selection to prevail over counteracting individual selection. *Evolution* **36,** 832–842 (1982).
69. Long, J. C. The allelic correlation structure of Gainj and Kalam speaking peoples and interpretation of Wright's *f*-statistics. *Genetics* **112,** 629–647 (1986).
70. Kelly, R. C. *The Nuer Conquest: The Structure and Development of an Expansionist System* (Univ. Michigan Press, Ann Arbor, 1985).
71. Bowles, S., Choi, J.-K. & Hopfensitz, A. The co-evolution of individual behaviours and social institutions. *J. Theor. Biol.* (in the press).
72. Henrich, J. & Boyd, R. Why people punish defectors—weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *J. Theor. Biol.* **208,** 79–89 (2001).
73. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100,** 3531–3535 (2003).
74. Gintis, H. The hitchhiker's guide to altruism: Gene-culture co-evolution and the internalization of norms. *J. Theor. Biol.* **220,** 407–418 (2003).
75. Schotter, A. Decision making with naive advice. *Am. Econ. Rev.* **93,** 196–201 (2003).
76. Chaudhuri, A. & Graziano, S. *Evolution of Conventions in an Experimental Public Goods Game with Private and Public Knowledge of Advice* (Working Paper, Department of Economics, Univ. Auckland, 2003).
77. Harbaugh, W. T., Krause, K. & Liday, S. *Children's Bargaining Behavior: Differences by Age, Gender, and Height* (Working Paper, Department of Economics, Univ. Oregon, 2000).
78. Jorgensen, J. G. *Western Indians: Comparative Environments, Languages, and Cultures of 172 Western American Indian Tribes* (W. H. Freeman, San Francisco, 1980).
79. Otterbein, K. F. *The Evolution of War: A Cross-Cultural Study* (Human Relations Area Files Press, New Haven, 1985).
80. Soltis, J., Boyd, R. & Richerson, P. J. Can group-functional behaviors evolve by cultural-group selection—an empirical-test. *Curr. Anthropol.* **36,** 473–494 (1995).
81. Bornstein, G. & Ben-Yossef, M. Cooperation in intergroup and single-group social dilemmas. *J. Exp. Soc. Psychol.* **30,** 52–67 (1994).

Acknowledgements We gratefully acknowledge support by the Ludwig Boltzmann Institute for the Analysis of Economic Growth, by the Swiss National Science Foundation and by the MacArthur Foundation Network on Economic Environments and the Evolution of Individual Preferences and Social Norms. We thank G. Bornstein, S. Bowles, R. Boyd, M. Brewer, J. Carpenter, S. Gächter, H. Gintis, J. Henrich, K. Hill, M. Milinski, P. Richerson, A. Riedl, K. Sigmund, E. A. Smith, D. S. Wilson and T. Yamagichi for comments on the manuscript, and M. Naef, D. Reding and M. Jörg for their research assistance.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.F. (efehr@iew.unizh.ch).

© 2003 Nature Publishing Group

Copyright © 2003 EBSCO Publishing