Analyse & Kritik 27/2005 (© Lucius & Lucius, Stuttgart) 113–135

Terence C. Burnham/Dominic D. P. Johnson

The Biological and Evolutionary Logic of Human Cooperation

Abstract: Human cooperation is held to be an evolutionary puzzle because people voluntarily engage in costly cooperation, and costly punishment of non-cooperators, even among anonymous strangers they will never meet again. The costs of such cooperation cannot be recovered through kin-selection, reciprocal altruism, indirect reciprocity, or costly signaling. A number of recent authors label this behavior 'strong reciprocity', and argue that it is: (a) a newly documented aspect of human nature, (b) adaptive, and (c) evolved by group selection. We argue exactly the opposite; that the phenomenon is: (a) not new, (b) maladaptive, and (c) evolved by individual selection. In our perspective, the apparent puzzle disappears to reveal a biological and evolutionary logic to human cooperation. Group selection may play a role in theory, but it is neither necessary nor sufficient to explain human cooperation. Our alternative solution is simpler, makes fewer assumptions, and is more parsimonious with the empirical data.

1. Introduction

In recent years, a group of scholars has advanced a set of heterodox claims about the origin and nature of human cooperation in a burst of papers published in leading journals (Bowles/Gintis 2002; Boyd et al. 2003; Fehr/Fischbacher 2003; 2004; Fehr/Fischbacher/Gächter 2002; Fehr/Gächter 2002; Fehr/Henrich 2003; Fehr/Rockenbach 2003; Gintis 2000; Gintis et al. 2003). They write that "human prosocial behavior ... is fundamentally incompatible with the economist's model of the self-interested actor and the biologist's model of the self-regarding reciprocal altruist" (Gintis et al. 2003, 169). The solution to this puzzle, the group of scholars (hereafter 'the Collective') argue, is an altruistic human predisposition to work for the good of the group, arising by group selection.

This view contradicts decades of work in biology, economics and other fields, which explains puzzling behavior as merely *apparent* altruism that just disguises an underlying pursuit of self-interest via kin selection, reciprocal altruism, indirect reciprocity, or costly signaling (Alexander 1979; 1987; Hamilton 1964; Nowak/Sigmund 1998a;b; Trivers 1971; 1985; Zahavi 1975; Johnson/Stopka/Knights 2003).

Resolving this debate is of immense importance. Quite apart from understanding the origins of cooperation and sociality, and why and when humans co-

operate today, it is essential for designing social contracts and institutions that will foster cooperation and facilitate conflict resolution. Despite its importance, there have been no reviews of this work by anyone other than the Collective themselves (see, however, Price/Cosmides/Tooby 2002 for an alternate explanation of the empirical evidence). Here we examine each of their claims in turn, and present an alternative explanation for apparently altruistic behavior.

2. What is Strong Reciprocity?

Strong reciprocity (hereafter SR) is a term created by the Collective. It is a *descriptive* term for the phenomenon that "people tend to behave prosocially and punish antisocial behavior at cost to themselves, even when the probability of future interactions is low or zero. We call this *strong reciprocity.* (Gintis 2000, 177) The term is actually a confusing misnomer—it describes behaviors, such as cooperation in one-shot prisoner's dilemmas, that do not involve (or even allow for) any reciprocity. However, the key point here is that we do not contest the existence of the phenomenon itself. People really do cooperate and punish at a cost to themselves and few, if any, scholars dispute that. The question of interest is not *if* but *why.*

The study of voluntary and costly behavior dates back to Darwin, who asked in *The Origin of Species*, "Can we consider the sting of the wasp or of the bee as perfect which ... inevitably causes the death of the insect"? (Darwin 1859, ch. 6) Darwin fretted that suicidal stings contradicted his theory because natural selection should eradicate behaviors that are voluntary and costly to the individual. A century and a half later, four main non-altruistic explanations for individually costly behavior have emerged.

Kin selection (Hamilton 1964) explains costly acts as benefiting genetic relatives (such as Darwin's bees). Among genetically *unrelated* individuals, reciprocal altruism (Trivers 1971), indirect reciprocity (Alexander 1979; 1987; Nowak/Sigmund 1998 a;b), and signaling (Gintis/Smith/Bowles 2001; Zahavi 1975) explain costly acts as part of a longer-term strategy that in fact advances individual interests overall. Via these four mechanisms, many behaviors that appear costly are beneficial because they bring repayment to the individual or to genetic relatives.

By contrast, the Collective advocate a 'genuine' altruistic force—not explained by these four mechanisms—in human cooperation. Cooperative acts subject to any of the four non-altruistic mechanisms are labeled as 'weak reciprocity'. Any residual cooperation—cooperation where costs cannot be repaid via kin-selection, reciprocal altruism, indirect reciprocity, or costly signaling—is labeled 'strong reciprocity' (see Table 1).

Our disagreement with the Collective centers on the ultimate cause of a particular set of altruistic behaviors—the behaviors themselves are neither controversial nor disputed. The standard definition of altruism is an action that includes three necessary and sufficient conditions: i) it is voluntary, ii) costly to the actor, and iii) benefits one (or more) organisms.

	Puzzling behavior	Direct effect on actor (immediate payoff)	Indirect effect on actor (payoffs in the future)	Net effect on actor
'Strong Reciprocity' (All costly cooperation and/or punishment that cannot be repaid at the individual level)	• Costly cooperation • Costly punishment of non-cooperators	**Negative**	**Zero** (or not positive enough to compensate for costs)	**Negative** (Puzzle is *not* resolved at the level of the individual, since apparently costly behavior *does not* redound to benefit of the individual)
'Weak Reciprocity' (Reciprocal altruism, indirect reciprocity, and costly signaling)	• Costly cooperation • Costly punishment of non-cooperators	**Negative**	**Positive**	**Positive** (Puzzle is resolved at the level of the individual, since apparently costly behavior redounds to benefit of the individual)

Table 1: 'Strong reciprocity' (in dark grey) and 'weak reciprocity' (in light grey) among non-relatives as defined in the literature by the Collective.

Both we and the Collective agree that humans in and outside of laboratories behave altruistically. Furthermore, we and the Collective agree on the theoretically possible evolutionary sources of altruism. They are kin selection, reciprocal altruism, indirect reciprocity, costly signaling and group selection. Where we differ is on the origin of the mechanisms that create the behavior that the Collective label strong reciprocity. The Collective argue that the existing data provide support for group selection as a source for human altruism. We do not believe that the existing data provide support for group selection. We argue that the existing data are more parsimoniously explained as the artifact of individually selected mechanisms operating in particular evolutionarily novel settings.

Thus, we differ on how to interpret behavior that is good for the group, but bad for the individual. Is it: i) an artifact of individually selected mechanisms, or ii) produced by group-selected mechanisms? We are not aware of any existing terminology that differentiates these two sorts of behavior. Accordingly, we introduce the term "genuine" altruism for the latter.

The Collective have performed numerous experiments showing that people do take costly cooperative actions even when they cannot recoup those costs. All of these demonstrations share three features: experimental subjects are not genetic relatives (ruling out benefits to kin); subjects never interact with each other again (ruling out direct repayment); and subjects' decisions are anonymous (ruling out reputation formation or signaling).

3. Examples of Strong Reciprocity

Strong reciprocity (SR) comes in two forms: *positive* SR (costly cooperation); and *negative* SR (costly punishment of non-cooperators). Below, we summarize one empirical demonstration of each form (they are reviewed extensively elsewhere (Fehr/Fischbacher 2003; Gintis et al. 2003)).

3.1 Positive Strong Reciprocity

In a 'trust game' (Fehr/Gächter/Kirchsteiger 1997), workers form contracts with firms and then choose how hard to work (both the 'workers' and the 'firms' are laboratory subjects). Workers who toil harder than the minimum feasible amount incur costs, but create extra profits for the firm. Effort levels are not enforceable, so after agreeing to an employment contract, workers can be 'lazy' and still receive full pay. The benefits of work (to the firm) exceed the cost (borne by the workers) so worker effort increases the combined payoff of the two parties.

If people acted only to maximize monetary payoffs, then the prediction in this setting would be lazy workers and stingy bosses. However, empirical results reveal many hard workers and many generous bosses. Workers who work harder than required—without the possibility of additional payment—exhibit positive SR.

3.2 Negative Strong Reciprocity

'Altruistic Punishment' (Fehr/Gächter 2002) is an example of negative SR, revealed in a public goods game modified to allow players to sanction each other. In the standard public goods setting, players choose how much to contribute to a public account. Contributions are multiplied and then divided equally among all the players, regardless of their level of contributions. Although the maximum possible earnings come when all cooperate, each individual has a dominant money-maximizing strategy to contribute nothing (thus, positive contributions represent positive SR, and in a wide-range of studies, whether interactions are repeated or not, subjects indeed contribute between 30 and 70 percent of their resources to the public account) (Ledyard 1995).

Negative SR occurs when players take money out of their own pocket to punish non-cooperators. This happens when players are allowed to see the public goods contributions of other players and inflict monetary punishments on them (players' actions are visible, but their identities remain hidden). Subjects voluntarily administer punishment, which is altruistic because it: (a) is costly to the punisher; and (b) benefits others.

Experimental design guarantees that the act of punishing decreases the punisher's payoff. Because the punishment phase is final and anonymous, there can be no indirect benefits that accrue to the punisher—hence, negative SR.

We agree with the Collective that humans attempt to cooperate with anonymous strangers whom they will never meet again. However, we challenge their interpretation that this phenomenon represents genuine altruism (as defined above). Rather than rely upon group selection as a cause of this behavior, we offer a simpler, more parsimonious alternative.

4. Strong Reciprocity is Not New

According to the Collective, costly cooperation constitutes "new knowledge" revealed by a slew of "recent experimental research" (Gintis et al. 2003, 153, 169). To the contrary, Table 2 summarizes previous research documenting precisely the same phenomenon. In each case, the seminal studies predate the work of the Collective by years or decades.

Furthermore, the earlier studies focused on the same issue—costly cooperation that was not repaid. Cooperation in prisoner's dilemma, for example, spawned a vast literature largely because people cooperated more than was expected (Axelrod 1984; Kagel/Roth 1995; Poundstone 1992). Similarly, the 1982 'ultimatum game' study (Güth/Schmittberger/Schwarze 1982) became famous because subjects incurred punishment costs that cannot be recouped.

Even in a public goods game setting, Yamagishi's 1986 study (Yamagishi 1986) found that, given the opportunity, individuals voluntarily engaged in costly punishment. There is one relevant difference between Yamagishi's 1986 work and the later work of the Collective. Yamagishi has stable groups that interact over multiple rounds. In a repeated game setting, only punishment in the final round

Phenomenon	Dates and Setting	Voluntary and costly behavior	Direct effect on actor (immediate payoff)	Indirect effect on actor (payoffs in the future)	Direct effect on others
Strong Positive Reciprocity (costly cooperation)	Contract enforcement (Fehr, Gachter, and Kirchsteiger 1997)	Voluntary gift in the form of extra worker effort.	**Negative** Gift decreases earnings	**None** Gift is anonymous and occurs in a final interaction	**Positive** Gift increases the payoff to the counterpart
	Trust Games (Berg, Dickhaut, and McCabe 1995)	Voluntary gift in the form of cash.	**Negative** Gift decreases earnings	**None** Gift is anonymous and occurs in a final interaction	**Positive** Gift increases the payoff to the counterpart
	Prisoner's Dilemma (various, see Axelrod 1984; Kagel and Roth 1995; Poundstone 1992)	Unilateral cooperation where defection would earn more	**Negative** Cooperation earns less than defection	**None** Cooperation is anonymous and occurs in a final interaction	**Positive** Another party benefits from cooperation
Strong Negative Reciprocity (costly punishment of non-cooperators)	Public goods game with punishment (Fehr and Gachter 2002)	Costly punishment in final interaction after a public goods decision	**Negative** Punishment is costly to perform	**None** Punishment is anonymous and occurs in a final interaction	**Negative** Being punished is costly
	Public goods game with punishment (Yamagishi 1986)	Costly punishment in final interaction after a public goods decision	**Negative** Punishment is costly to administer	**None** Punishment is anonymous and occurs in a final interaction.	**Negative** Being punished is costly
	One-shot ultimatum game (Guth, Schmittberger, and Schwarze 1982)	Irrevocable rejection of a positive offer in a bargaining game	**Negative** Rejection earns less than acceptance	**None** Rejection is anonymous and occurs in a final interaction	**Negative** Being rejected is costly

Table 2: Strong reciprocity is not a new phenomenon. dark grey denotes recent demonstrations by the Collective; light grey denotes earlier demonstrations by others.

is clearly non-strategic. So Yamagishi's subjects had only one chance to demonstrate punishment that does not redound to their advantage whereas subjects in the Collective's setting have more than one opportunity. Subjects in both experiments do use their punishment technology even in terminal interactions.

Yamagishi's investigation of punishment technology is in one respect more nuanced than that of the Collective. The altruistic punishment study of the Collective used a fixed ratio of 3:1 for the penalty to the punished for every dollar invested in punishment. Yamagishi's work includes ratios of 2:1 and 1:1. Subjects in both Yamagishi treatments used the punishment technology. However, the Collective's finding that cooperation increases over time is only apparent in Yamagishi's treatments with the higher ratio. Thus, the Collective's empirical result (Fehr/Gächter 2002) is in some ways a special case of Yamagishi's 1986 paper.

The key features of what the Collective label altruistic punishment were demonstrated in a public goods game setting and published more than a decade before the Collective began work in the area.

Experiments by Fehr and others are indeed the clearest demonstrations yet that cooperation and punishment occur even when they are costly and voluntary for the actor. What is striking, however, is that older experimenters saw no puzzle, despite drawing equivalent conclusions about human behavior (that people cooperate and punish 'too much'). Once the Collective conducted experiments that explicitly ruled out each possible selfish incentive one by one (the lurking alternative explanations), this seems to have generated a spurious puzzle by assuming that the removal of apparent incentives had also perfectly ruled out human *responses* to those supposedly missing incentives (e.g. that 'anonymous' experiments conducted in the laboratory were really perceived as anonymous by the participants). If selfish incentives were absent, but cautionary selfish concerns still present, this would generate the spurious puzzle. Hence the emergence of revisionist explanations that were required to account for the 'new' findings.

In addition to these earlier *empirical* demonstrations of costly cooperation, most *theoretical* developments surrounding SR are also predated by Robert Trivers' famous 1971 article on reciprocal altruism (Trivers 1971).

Trivers noted a deep-rooted human disposition to act altruistically, even towards unrelated others and beyond that expected of a rational actor (Trivers 1971). Reciprocal altruists start off being 'nice' (positive SR), but they also punish those who attempt to exploit the system (negative SR), a trend corroborated in Axelrod's work on 'tit-for-tat' strategies (Axelrod 1984). Trivers wrote:

> "Once strong positive emotions have evolved to motivate altruistic behavior, the altruist is in a vulnerable position because cheaters will be selected to take advantage of the altruist's positive emotions. This in turn sets up a selection pressure for a protective mechanism. Moralistic aggression and indignation in humans was selected
>
> (a) to counteract the tendency of the altruist, in the absence of any

> reciprocity, to continue to perform altruistic acts for his own emotional rewards
>
> (b) to educate the unreciprocating individual by frightening him with immediate harm or with future harm of no more aid." (Trivers 1971, 49)

Trivers also applied reciprocal altruism to larger groups:

> "Selection may favor a multiparty altruistic system in which altruistic acts are dispensed freely among more than two individuals, an individual being perceived to cheat if in an altruistic situation he dispenses less benefit for the same cost than would others, punishment coming not only from the other individual, but from the others in the system." (Trivers 1971, 52)

The 'multi-party' applications of reciprocal altruism (crucial to explaining cooperation among human groups and societies at large) were later expanded and formalized by others, notably Richard Alexander, Martin Nowak and Karl Sigmund (Alexander 1987; Nowak/Sigmund 1998 a;b). Trivers' theory is therefore applicable to larger groups, despite the Collective's concern that: "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." (Fehr/Fischbacher 2003, 788) Reciprocity breaks down when groups become large (Boyd/Richerson 1988), but nevertheless remains an important source of cooperation among small hunter-gatherer groups (in which spatial structure, status hierarchies and social networks anchor repeated cooperation with specific neighbors and acquaintances).

In summary, SR shares key elements with both previous empirical and theoretical work on cooperation—it is a well-documented, multi-party, 'be nice, but punish' system fostering cooperation. Therefore, the unique aspect of 'strong reciprocity' is to view costly cooperation—in settings structured to make repayment impossible—as adaptive and genuinely altruistic. In the next section, we show that *even if the Collective were right* about the origins of SR, its expression in experiments is *necessarily maladaptive*.

5. Strong Reciprocity is Maladaptive

In a book chapter devoted to "discuss the evidence bearing on the question of whether strong reciprocity represents adaptive or maladaptive behavior", the Collective conclude, "the evidence suggests that it [SR] is adaptive" (Fehr/Henrich 2003, 55). Elsewhere they write, "Some behavioral scientists have suggested that the behavior we have described in this article [SR] is maladaptive ... but we do not believe that this critique is correct" (Gintis et al. 2003, 168). However, we show that applying appropriate definitions inevitably leads to the conclusion that SR must be maladaptive.

Following Wilson (1975), we define adaptation initially as "any structure, physiological process, or behavioral pattern that makes an organism more fit to survive and to reproduce in comparison with other members of the same species" (Wilson 1975, 577).

Using this definition, traits are either advantageous or disadvantageous depending on their fitness effects. While such a binary definition works for some traits, it is critical to separate the fitness effects of a trait in two different time periods. The first is the period in which the trait arose, and the second is in the present time (Gould/Vrba 1982), allowing a "crisp dissection of a key problem in evolutionary biology—the distinction between historical origin and current utility" (Gould 2002, 1216). This richer framework transforms the one-dimensional view of adaptation (simply 'advantageous' or 'disadvantageous') into two dimensions, resulting in four types of trait (see Figure 1).

		Fitness effect of trait at point of historical origin	
		Fitness increase	**No fitness increase**
Fitness effect of trait in modern environment	**Fitness increase**	ADAPTIVE Strong Reciprocity (according to the Collective)	EXAPTIVE
	No fitness increase	MALADAPTIVE Strong Reciprocity (in reality)	SIDE EFFECT

Figure 1: Strong reciprocity is maladaptive. Dark grey denotes the Collective's classification of strong reciprocity, light grey denotes its actual classification. If strong reciprocity evolved by group selection (as the Collective claim) it is located on the left-hand column of the chart (traits that increased fitness in our evolutionary history), and since strong reciprocity is costly in modern settings, it must be located in the lower row. 'Side effects' are incidental traits that confer fitness neither now nor in the past. 'Exaptive' traits are those that provide fitness benefits today, but were co-opted from some other origin (the classic example is wings, thought to originate from cooling devices) (Gould/Vrba 1982).

Costly behavior should be labeled as maladaptive if it fits two criteria: (a) the behavior is produced by a physiological/psychological system that was shaped by natural selection to increase fitness; and (b) the behavior *in the environment under consideration* does not increase fitness. As an example, Irons argues that while the human craving for salt, fats and sweets increased fitness in ancestral environments, those same preferences are maladaptive when we forage in modern supermarkets or junk food eateries (Irons 1998).

The Collective claim SR is adaptive and it evolved by group selection. Yet, these two claims are not consistent. Even if SR did arise by group selection, it would be maladaptive in its laboratory manifestations. Consider the two group selection models of SR (Boyd et al. 2003; Gintis 2000). In multiple, competing groups, individual 'strong reciprocators' who sacrifice for the good of the group have lower returns than group mates who refrain from such altruism, and therefore those who exhibit SR shrink as a percentage of the group (given at least one non-altruist). However, in spite of losing every within-group competition, SR can persist if groups with more altruists outperform competing groups with fewer altruists by a sufficiently wide margin (Wilson/Sober 1994). Therefore, a minimum condition to allow SR to persist is that its altruistic effects must redound to the benefit of the group.

In the laboratory manifestations of SR, of course, the consequences of altruistic actions are conferred on unknown and one-off laboratory subjects. Thus, no relevant 'group' can benefit. As an analogy to clarify this point, consider applying the definition of maladaptive to kin selection: If SR were a behavioral tendency that arose by *kin selection*, but was expressed in a laboratory context that benefited *people who were not genetic relatives*, then the experimental manifestation would be maladaptive. By the same logic, if SR were a behavioral tendency that arose by *group selection*, but was expressed in a laboratory context that benefited *people who were not in one's group*, then the experimental manifestation would again be maladaptive.

Group selection is therefore no more consistent with SR than is kin selection. The Collective exclude kin selection as the ultimate cause of SR because laboratory subjects are not genetically related. Exactly the same logic excludes group selection—laboratory subjects are not drawn from the same evolutionarily relevant group either. They may behave *as if* they were (people often do identify with arbitrary experimental groups (Tajfel 1974)), but in that case we may just as well assume that anonymous subjects behave *as if* they are related (and invoke kin-selection), *as if* they are destined to meet again (and invoke reciprocal altruism), or *as if* they are observed by others (and invoke indirect reciprocity). On this point, group selection is no better at explaining cooperation than any of the traditional theories.

Hence, SR is maladaptive precisely because experimental subjects who exhibit SR help people who are not in any evolutionarily relevant group, just as they are not kin. Subjects are drawn from large urban populations and placed into temporary experimental groups of a handful of individuals. These subjects leave these temporary groups behind immediately after the experiment. Those who sacrifice do so for the good of anonymous strangers whom they will never

see again; thus, the altruism of SR does not (and cannot) redound to members of any relevant group subject to the forces of selection.

The group selection models of SR do not provide an adaptive explanation for the laboratory manifestations of SR. They do not even apply in most cases because they deal with specific games: a 2-stage game with punishment in one model (Boyd et al. 2003) and a public goods game in the other (Gintis 2000). Further scrutiny reveals that these models do not address *any* of the empirical manifestations of SR.

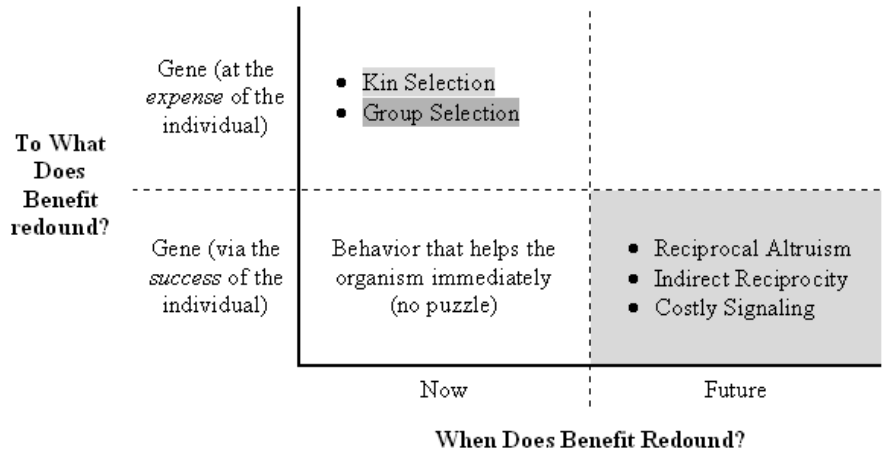

Figure 2: Solutions to the puzzle of cooperation. Dark grey denotes the Collective's favored solution, light grey denotes ours. Voluntary and costly acts can increase fitness via two conceptually independent routes: First, kin selection (Hamilton 1964) and group selection in its modern formulation (Wilson/Sober 1994) are routes to immediate *genetic level* advantage—even though the organism itself may experience a net loss (upper left quadrant). Second, reciprocal altruism (Trivers 1971), indirect reciprocity (Alexander 1979; 1987; Nowak/Sigmund 1998a;b), and costly signaling (Zahavi 1975) are routes to *organism level* advantage - despite an immediate 'direct' loss, there is a net gain overall due to 'indirect' repayment by others in subsequent periods (lower right quadrant).

The first model asks, "When can cooperation by sustained?" and finds "the condition for cooperation is $c < \delta_* \pi$", where c is the cost of SR, δ_* is the probability that "the group will persist ... provided that members cooperate", and π is an individual's total expected fitness when he and all other group members contribute (Gintis 2000, 173–4). In the experimental manifestations of SR the probability that the transient group will persist is zero, thus, for cooperation to be sustained the cost of SR, c, must be less than zero. However, by definition, the cost of SR is greater than zero (cooperation that is not costly is 'weak reciprocity'). Ergo, cooperation in this model cannot be sustained within transient groups.

The second model concludes, "In this model, group selection leads to the evolution of cooperation only if migration is sufficiently limited to sustain substantial between-group differences in the frequency of defectors" (Boyd et al. 2003). In the experimental manifestations of SR, the groups dissolve after one interaction meaning that the migration rate is 100%.

Thus, neither of these models of SR provides an adaptive explanation of SR among subjects drawn from the enormous populations of cities like Zurich or New York. Indeed, one of the group selection models of SR (Boyd et al. 2003) finds cooperation is limited to groups of less than a few hundred individuals, and breaks down above that. Ancestral humans may have lived in such small groups, but the subjects in the experimental demonstrations of SR certainly do not.

Thus, one has to conclude that if SR arose by natural selection (irrespective of whether at the individual or group level), then it is maladaptive in laboratory experiments. Voluntary and costly behaviors are adaptive if sufficient benefits accrue to (a) the altruists' genes through kin or group selection, or (b) to the individual in the future (see Figure 2). None of these mechanisms are employed adaptively in the empirical demonstrations of SR, since subjects cannot be repaid, and interact with people who are neither kin nor group mates.

The Collective dismiss the idea that SR "results from the maladaptive operation of a psychology that evolved in ancestral environments" (Fehr/Henrich 2003, 78). Ironically, we can now see that this exact critique applies to their group selection theories as well. If *the Collective are right* about the ultimate cause of SR—group selection, *they are wrong* about its manifestation in their experiments. Alternatively, they may be wrong about both.

6. Strong Reciprocity Arose by Individual Selection

The Collective argue that SR represents a genuine individual sacrifice for the good of group. However, none of their arguments provide support for group selection. Rather, existing evidence suggests that SR arose by individual selection, and therefore is not genuinely altruistic.

To understand behavior, especially maladaptive behavior like SR, it is vital to distinguish between the 'proximate' cause (the physiological mechanism) and the 'ultimate' cause (the evolutionary 'goal') of the behavior in question (Mayr

1961; Tinbergen 1963; 1968). Focusing only on one without consideration for the other will often lead to confusion.

As an example, herring gulls will preferentially care for artificial, over-sized eggs while their offspring in real eggs die nearby. At first glance this appears puzzling. But researchers discovered that in natural settings, a 'bigger is better' rule (the proximate cause of behavior) works to increase fitness (the ultimate cause) by favoring those eggs most likely to produce the best offspring—the bigger ones (Baerends/Drent 1982 a;b; Baerends/Krujit 1973). In the initial experiment, the maladaptive behavior of favoring experimental eggs is the product of the usually beneficial mechanism (favoring larger eggs over smaller) operating within an evolutionarily novel environment (of human altered eggs) contrived precisely to fool these mechanisms.

Similarly, human cooperative tendencies have proximate causes, and the distinction between proximate and ultimate causation means that such tendencies—even if they arose only by individual selection—will sometimes fail to achieve their evolutionary goals in some evolutionarily novel settings, such as laboratory experiments. We may fool our subjects but this need not fool our evolutionary logic also.

To see why, let's recall the puzzle: 'strong reciprocity' (voluntary and costly cooperative act that is not repaid at the individual level) occurs in experiments specifically constructed to make repayment impossible. Now, unless mechanisms to exploit reciprocal altruism, indirect reciprocity, or signaling work *so perfectly as to produce zero cooperation in such anonymous and final interactions*, people will sometimes cooperate even when they cannot be repaid. A number of cues may trigger such cooperative behavior. Such behavior will be classified as SR, yet it does not reveal genuine altruism (just as herring gulls are not trying to kill their own offspring). Remember that SR is maladaptive regardless of whether it originated via individual or group selection. Therefore, the real competing explanations for SR are: (a) the maladaptive expression of individually selected mechanisms; or (b) the maladaptive expression of group selected mechanisms. The former is commonly observed (e.g. eating junk food), but the latter is speculative. Those who favor group selection as an important force guiding human behavior must address the daunting pre-requisites regarding within-group stability, among-group variation, migration, extinction, and countervailing action of within-group selection for self-interest (Wilson/Sober 1994).

To argue that group selection is the cause of SR, the Collective need to produce compelling evidence, *consistent* with group selection but *inconsistent* with individual selection. In the absence of such data, we show below that the more parsimonious explanation is that maladaptive SR is the product of individual selection.

6.1 Three Flawed Arguments

Presently, the Collective do not provide any evidence that suggests group selection supercedes individual selection. The three central arguments they present are flawed.

The Collective's first argument attempts to lower the bar for group selection by prematurely declaring that SR "cannot be explained in terms of self-interest" (Gintis et al. 2003, 154). With individual selection supposedly ruled out, the mere mathematical possibility of group selection must, the Collective argue, be considered as the solution to the puzzle of SR (Fehr/Fischbacher 2003).

This argument suffers from two errors. First, SR *can* be explained in terms of self-interest (as explained above). Second, the Collective confuse the valid competing hypotheses by incorrectly suggesting that group selection provides an *adaptive* explanation for SR. As shown above, it does not.

The Collective's second type of argument is ineffective because it presents evidence that is consistent with both group selection and individual selection. For example, they write, "what is the evidence for cultural group selection? There is quite strong evidence that group conflict and warfare were widespread in foraging societies." (Fehr/Fischbacher 2003, 790) Yet, the existence of inter-group conflict is of course perfectly consistent with individual selection as well (Keeley 1996; LeBlanc/Register 2003; Wrangham 1999). Similarly, the Collective cite the effort that parents take to teach children social skills (Fehr/Fischbacher 2003). Obviously, there is no need to invoke group selection to explain helping behavior among close kin. Nor does evidence of cultural variation in SR contradict individual selection (Henrich et al. 2001, see also Price 2005). Phenotypic plasticity is common in species of all Earth's taxa. We repeat: group selection could be involved in strong reciprocity, but it is not necessary to explain the phenomenon.

The Collective's third type of argument, paradoxically, uses evidence *of individual selection* to argue in favor of group selection. For example, they argue that, "(t)he experimental evidence unambiguously shows that subjects cooperate more in two-person interactions if future interactions are more likely" (Fehr/Fischbacher 2003, 788). Such studies are evidence that individual selection has built people to care about their reputations.

In interpreting these data, however, the Collective write that individual selection can explain SR *only if* human behavioral rules "do not distinguish between kin and non-kin" (in the case of kin-selection), "did not depend on the probability of future interactions with potential opponents" (for reciprocal altruism), or "did not depend on our actions being observed by others" (for indirect reciprocity and costly signaling) (Fehr/Gächter 2003, 912). This is false. The fact that individual selection has built sophisticated mechanisms to manage reciprocity, reputation or signaling in no way implies that those mechanisms are inflexible, or will be triggered perfectly in different contexts. Mechanisms to modulate cooperation can be both sophisticated and imperfect—adjusting to the likelihood of future interactions and yet still sometimes cooperating 'too much' in anonymous and final interactions.

The Collective are essentially debunking a straw man model of individual selection, in which behavioral rules are seen as fixed (see Figure 3). The correct model (proximate causation based in individual selection) predicts that cooperation *will vary* to some extent based on the *perceived* probability of forming reputations. (It is this straw-man that makes the Collective's preoccupation with the supposed frequency and costliness of one-shot interactions in our evolutionary past entirely irrelevant to the debate—such incidents are merely a further source of selection for mechanisms that are responsive to the varying potential for reputation formation.)

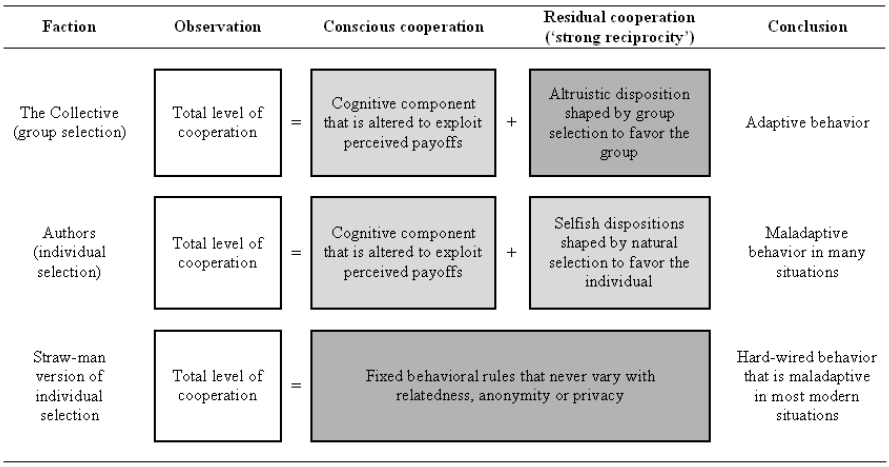

Figure 3: Models of cooperation. Light grey indicates behavior *expected to vary* with relatedness (the potential for kin-selection), anonymity (the potential for reciprocity), or privacy (the potential for reputation formation); dark grey indicates behavior *independent* of relatedness, anonymity or privacy. Once the incentives for *conscious* cooperation are experimentally excluded, only the evolutionary legacy hypothesis can account for why the remaining *residual* cooperation ('strong reciprocity') also varies with apparent but inconsequential relatedness, anonymity and privacy.

Human cooperation is modulated both consciously and subconsciously. Human brains are obviously capable of adjusting their conscious level of cooperation to reap apparent and available rewards (the first component of Figure 3). However, once these rewards to cooperation are plainly denied (by careful experimentation 'ruling out' such incentives), people continue to cooperate—the

phenomenon of SR. This residual cooperation represents subconscious mechanisms (the second component of Figure 3).

The key question, therefore, becomes which theory—maladaptive individual selection or maladaptive group selection—should be credited for the 'residual' cooperation that is SR (Figure 3). The answer requires careful experiments to tease apart the causal mechanisms. Fortunately, there are differences in the predictions of these competing approaches. Individually selected mechanisms predict that even residual cooperation will vary with perceived anonymity or expectations of future interactions (due to proximate cues for reputation management), whereas cooperation based on group selection does not predict any such relationship.

One study that does differentiate the competing hypotheses varied the level of anonymity and measured the effect on residual cooperation (Rege/Telle 2004). In a public goods game played with no possibility of future interactions, subjects were drawn from a large urban population, screened to ensure that they had never met each other before, and forced to leave the lab separately with several minute delays. After all decisions were made, some of the subjects had to reveal their decision publicly (these subjects knew this before they made their decisions).

The results betray individual selection as an important cause of SR: subjects who knew they would be identified with their decisions contributed 64% more (72% in the public case, 44% in the non-public case) than those who played anonymously—even though all cooperation in this experiment qualifies as SR (the subjects do not face any monetary rewards or punishment for their behavior). Hence, the level of SR itself is powerfully changed by a parameter specifically predicted by individual selection, and not predicted by group selection. If altruists working for the good of the group cause SR, why would they care about the privacy of their actions?

This one experiment does not, of course, prove that there is no role for group selection in SR. But it does support individual selection, as well as suggest the type of work that the Collective ought be doing: those who favor group selection should manipulate parameters unique to group selection and test whether those parameters alter the level of SR. They have rejected individual selection theories at a (false) theoretical level, and not at an empirical level.

Behavior tends to be based on specific environmental cues. Human cooperative mechanisms may interpret such cues to suggest that reciprocity or reputation is possible even when it is experimentally 'ruled out' and 'impossible'. If you ask experimental subjects, they may say that they 'know' repayment is impossible. Even so, their behavioral mechanisms may be influenced by emotional rewards to cooperate even when there is no chance for individual gain. It does not mean that the mechanisms were built by group selection (even the slightest possibility of exposure of one's selfish actions may be enough to motivate cooperation—a 'better-safe-than-sorry' heuristic). In fact, even a bias to cooperate too much may be a better error than cooperating too little, if one is sometimes watched when thought to be alone, or one's actions are discovered after the event.

While group selection remains possible and plausible, as an explanation of

SR it is neither necessary (since individual selection is more parsimonious) nor sufficient (since it cannot explain why SR varies with anonymity or reputation).

Stephen Jay Gould and Richard Lewontin have long warned against the kind of 'adaptationist' logic evident in the Collective's work: "One must not confuse the fact that a structure [or behavior] is used in some way ... with the primary evolutionary reason for its existence." (Gould/Lewontin 1979) The simple fact that SR sometimes confers benefits on others does nothing to support the Collective's assertion that those who exhibit SR are altruists working for the good of the group. Neither were the Herring gulls working for the good of the experimental eggs.

7. Conclusions

7.1 The Puzzle Vanishes

The phenomenon that humans incur costs to cooperate and punish among anonymous strangers is not new, is maladaptive, and seems to be caused by mechanisms that arose by individual selection. Hence, we see no puzzle in human cooperation.

Behavioral mechanisms are not perfect, always-optimal, goal seeking devices, but rather context-specific physiological systems that respond to environmental cues in order to engage what was, on average over the course of evolutionary history, the appropriate action. When those cues convey information out of context, then proximate mechanisms will often, unsurprisingly, produce maladaptive and costly behavior. Consequently, we see no need for the misnomer "strong reciprocity" to describe cooperative dispositions that are not repaid. To accept it would be to invite a host of similarly misleading labels for other ancestral mechanisms gone awry in modern settings, such as "obesity drive", "strong sperm bank cuckoldry", and "death instinct via adaptive heroin addiction".

Our view of cooperation produces testable hypotheses. Cues to anonymity and repeated interaction are predicted to alter the level of costly cooperation, even when those cues cannot alter payoffs. For example, in a recently completed study we found that—in anonymous and final interactions—subjects contributed significantly more to a public good when 'watched' by a robot with large, human-like eyes. The experiment was motivated by a hypothesis that human eyes would trigger non-conscious mechanisms that gauge privacy (Burnham/Hare 2006).

A second study found a positive correlation between human testosterone levels and punishing behavior (Burnham 2005). The study hypothesized that costly punishment in anonymous and final interactions is the maladaptive application of individually selected mechanisms to manage reputation. Because high testosterone men face lower costs to conflict in non-anonymous settings (Mazur/Booth 1998), it was predicted that they would be more likely to punish even when they could not be rewarded for their actions. A flurry of recent experiments similarly demonstrate that cooperation can be increased by activating cues of kinship (DeBruine 2002; Krupp/DeBruine/Barclay 2005), reciprocity (Price/Price/Curry 2005) and reputation (Haley/Fessler 2005).

We interpret these studies as supporting individual selection over group selection. A productive research agenda can draw upon this and other previous research to further elucidate the mechanisms that modulate cooperation. Each study needs careful development, but there are obvious candidates for further cues to anonymity and other factors affecting individual costs and benefits.

7.2 The Evolutionary Legacy Hypothesis

It is the biological basis of human cooperation that ensures the existence of costly cooperation. We therefore see a clear role for the evolutionary history of our species in this phenomenon. Accordingly, we set out here our own 'evolutionary legacy hypothesis', which can serve as a clear test and benchmark for future studies. This builds on a broad base of theory from the fields of ethology and behavioral biology that we have labored to set out explicitly here because it remains to be properly addressed by the Collective.

A variety of scholars suggest that systematic differences between ancestral and modern environments are the cause of many puzzling human behaviors (Bowlby 1969; 1973; Daly/Wilson 1983; Irons 1998; Tooby/Cosmides 1989; 1990; Wilson 1975; 1978). In a similar fashion, we believe that human cooperative tendencies may be explained, in part, as adaptive solutions to problems in ancestral environments (Bowlby 1969; 1973; Irons 1998). Because of the rapid pace of societal change, however, as well as ontological and phylogenetic constraints on evolution (Mayr 1961; Tinbergen 1963; 1968), human cooperative mechanisms are not in equilibrium with our environment.

As Robert Trivers suggests, our brains were shaped in a world that conferred net gains to those who granted initial generous outlays and punished cheats. This bias toward cooperation stems from a brain design selected over millions of years (Tooby/Cosmides 1990). Not surprisingly, therefore, it has persisted into the latest fraction of a percent of our history in which we find ourselves in cities, civilization and anonymous, one-shot laboratory experiments with strangers.

In short, we argue that there is a biological (proximate) and evolutionary (ultimate) logic to human cooperation. We predict that human cooperative mechanisms include design features that modulate behavior in non-anonymous and repeated environments, and that those mechanisms impact on the empirical level of cooperation even in anonymous and final interactions.

While our social environment has changed dramatically in the blink of a gene's eye, our brains have not, leaving humans with strange tendencies left over from a bygone era. Thus, the puzzle of strong reciprocity can be viewed as the result of a mismatch between ancient mechanisms and modern environments. This evolutionary view has already offered a novel perspective in some work in economics (Burnham 1997; 2003; Burnham/McCabe/Smith 2000; McCabe 2003; McCabe/Smith 2001; Smith 2003) and politics (Johnson 2004; Rosen 2004) and we expect it will become centrally important to a fuller understanding of human cooperation (Wilson 1998).

This view produces testable predictions that can differentiate between humans acting *as if* they were still in a world of kin, reciprocity and reputation

or *as if* they were group selected. Group selection predicts no covariance of SR (that is, the subconscious component only of human cooperation, see Figure 3) with anonymity, privacy, or kin. So any experiment that (a) removes all conscious *expectation* of future rewards, reputation or kin, *and* (b) finds that SR nevertheless varies with the subconscious *perception* of cues for these same three factors is, first, a direct falsification of the Collective's view and, second, support for ours.

7.3 Implications for Society

While we have focused on points of disagreement, we have significant common ground with the Collective. We admire their experimental work that elucidates the empirical patterns of human cooperation. Descriptive data on human cooperative behavior are central to understanding the phenomenon and to crafting appropriate institutions to promote it. However, its interpretation is crucial. If we misunderstand *why* humans cooperate, how can we encourage it?

Indeed, individual benefits still offer many advantages in modern life, despite the fact that in laboratory experiments they can be led astray. Even in a large modern city like New York, as far from our ancestral hunter-gatherer groups as we can imagine, the enormous group size and anonymity that would appear to rule out any cooperative incentives due to kin, reciprocity, or reputation mask the reality of human life: people interact within networks of family, friends and acquaintances, where kin, reciprocity and reputation are still crucial to everyday life and maintain the highest level of attention and social intrigue. For any one individual, New York still is a village—just many different villages overlapping each other. If we can identify the processes and cues that oil these cogs of social life then we may better construct the type of societies in which cooperation blooms. Individual selection still has a role to play in modern life, but group selection does not. Already, architects are moving away from residential tower blocks housing large groups of people that commonly led to social decay, and replacing them with model villages and communities that, essentially, bring out the best of our evolutionary roots.

Today, the divergence between proximate and ultimate causes of behavior is the source of much human strife. Proximate triggers continue to goad us into behavior that no longer fulfills an adaptive function. Fortunately, knowledge of such mismatch provides the normative basis for co-opting cooperative mechanisms towards the public good. Because mechanisms can be fooled, it is possible to design institutions in which individuals gain emotional rewards for helping society even at personal cost. Institutions, negotiations and markets must harness incentives that resonate with our true human nature, targeting specific stimuli that trigger cooperative dispositions, such as cues for reciprocity or reputation.

According to the Collective, "the moral sentiments that have led people to value freedom, equality, and representative government are predicated upon strong reciprocity" (Gintis et al. 2003, 154). This leads to the expectation that people will naturally act in a prosocial manner, a cozy assumption that we rely

on at our peril. History offers a mountain of warnings, from pirate rule in the 18^{th} century Caribbean to a host of past and present conflicts.

On the contrary, because humans cannot be relied upon to work for the good of the group, we must craft social, economic, environmental and political interactions to ensure cooperation against selfish temptation. If the human propensity to cooperate were shaped by group selection, why is punishment so essential to promote sacrifice for the group? It appears that punishment is necessary, paraphrasing Richard Sosis, "precisely because we are not likely to act for the benefit of the group when it is not in our own individual interests" (Sosis 2003, 140). Of course, this reality has been one of humankind's most fundamental intellectual and social challenges for centuries, from Plato and Adam Smith, to Marx and the Kyoto conference.

8. Acknowledgements

We would like to thank Jesse Bering, Lee Cronk, Ernst Fehr, Timothy Goldsmith, David Haig, Brian Hare, Marc Hauser, Jack Hirschleifer, Alex Kacelnik, Stephen Knights, Rob Kurzban, Matthew McIntyre, Jay Phelan, Michael Price, Vernon Smith, Pavel Stopka, Robert Trivers, Toshio Yamagishi, and Bill Zimmerman for comments and criticism.

Bibliography

Alexander, R. D. (1979), Natural Selection and Social Exchange, in: R. L. Burgess/T. L. Huston (eds.), *Social Exchange in Developing Relationships*, New York

— (1987), *The Biology of Moral Systems*, Hawthorne

Axelrod, R. (1984), *The Evolution of Cooperation*, New York

— /W. D. Hamilton (1981), The Evolution of Co-operation, in: *Science 211*, 1390–1396

Baerends, G. P./R. H. Drent (1982a), The Herring Gull and its Egg. Part I, in: *Behaviour, Suppl. 17*, 1–312

— / — (1982b), The Herring Gull and its Egg. Part II, in: *Behaviour 82*, 1–415

— /J. P. Krujit (1973), Stimulus Selection, in: R. A. Hinde/J. Stevenson-Hinde (eds.), *Constraints on Learning*, New York

Berg, J./J. Dickhaut/K. McCabe (1995), Trust, Reciprocity, and Social History, in: *Games and Economic Behavior 10(1)*, 122–142

Bowlby, J. (1969), *Attachment and Loss. Vol. I: Attachment*, New York

— (1973), *Attachment and Loss. Vol. II: Separation, Anxiety, and Anger*, New York

Bowles, S./H. Gintis (2002), Homo Reciprocans, in: *Nature 415*, 125–126

Boyd, R./H. Gintis/S. Bowles/P. J. Richerson (2003), The Evolution of Altruistic Punishment, in: *Proceedings of the National Academy of Sciences 100*, 3531–3535

— /P. J. Richerson (1988), The Evolution of Reciprocity in Sizable Groups, in: *Journal of Theoretical Biology 132*, 337–356

Burnham, T. C. (1997), *Essays on Genetic Evolution and Economics*, Parkland (dissertation.com)

— (2003), Engineering Altruism: A Theoretical and Experimental Investigation of Anonymity and Gift Giving, in: *Journal of Economic Behavior and Organization 50(1)*, 133–144

— (2005), Caveman Economics: Proximate and Ultimate Causes of Non-materially Maximizing Behavior, in: *Human Nature*

— /B. Hare (2006), Engineering Cooperation: Does Involuntary Neural Activation Increase Public Goods Contributions? In: *Human Nature*

— /K. McCabe/V. L. Smith (2000), Friend-or-Foe Priming in an Extensive Form Trust Game, in: *Journal of Economic Behavior and Organization 43(1)*, 57–73

Daly, M./M. Wilson (1983), *Sex, Evolution, and Behavior*, Belmont

Darwin, C. (1859), *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, London

DeBruine, L. (2002), Facial Resemblance Enhances Trust, in: *Proceedings of the Royal Society of London B 269*, 1307–1312

Fehr, E./U. Fischbacher (2003), The Nature of Altruism, in: *Nature 425*, 785–791

— / — (2004), Third-Party Punishment and Social Norms, in: *Evolution and Human Behavior 25*, 63–87

— / — /S. Gächter (2002), Strong Reciprocity, Human Cooperation and the Enforcement of Social Norms, in: *Human Nature 13*, 1–25

— /S. Gächter (2002), Altruistic Punishment in Humans, in: *Nature 415*, 137–140

— / — (2003), The Puzzle of Human Co-operation: A Reply, in: *Nature 421*, 912

— / — /G. Kirchsteiger (1997), Reciprocity as a Contract Enforcement Device: Experimental Evidenc, in: *Econometrica 65 (4)*, 833–860

— /J. Henrich (2003), Is Strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism, in: P. Hammerstein (ed.), *The Genetical and Cultural Evolution of Cooperation.*, Cambridge/MA

—/B. Rockenbach (2003), Detrimental Effects of Sanctions on Human Altruism, in: *Nature 422*, 137–140

Gintis, H. (2000), Strong Reciprocity and Human Sociality, in: *Journal of Theoretical Biology 206*, 169–179

— /S. Bowles/R. Boyd/E. Fehr (2003), Explaining Altruistic Behavior in Humans, in: *Evolution and Human Behavior 24*, 153–172

— /E. Smith/S. Bowles (2001), Costly Signaling and Cooperation, in: *Journal of Theoretical Biology 213 (1)*, 103–119

Gould, S. J. (2002), *The Structure of Evolutionary Theory*, Cambridge/MA

— /R. C. Lewontin (1979), The Spandrels of San Marco and the Panglossian Program: A Critique of the Adaptationist Programme, in: *Proceedings of the Royal Society of London 205*, 581–588

— /S. Vrba (1982), Exaptation—A Missing Term in the Science of Form, in: *Paleobiology 8*, 4–15

Güth, W./R. Schmittberger/B. Schwarze (1982), An Experimental Analysis of Ultimatum Bargaining, in: *Journal of Economic Behavior and Organization 3(4)*, 367–388

Haley, K/D. Fessler (2005), Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game, in: *Evolution and Human Behavior 26*, 245–256

Hamilton, W. D. (1964), The Genetical Evolution of Social Behavior I and II, in: *Journal of Theoretical Biology 7*, 1–16, 17–52

Henrich, J./R. Boyd/S. Bowles/C. Camerer/E. Fehr/H. Gintis/R. McElreath (2001), In Search of *Homo economicus*: Behavioral Experiments in 15 Small-Scale Societies, in: *American Economic Review 91(2)*, 73–78

Irons, W. (1998), Adaptively Relevant Environments Versus the Environment of Evolutionary Adaptedness, in: *Evolutionary Anthropology 6(6)*, 194–204

Johnson, D. D. P. (2004), *Overconfidence and War: The Havoc and Glory of Positive Illusions*, Cambridge/MA

— /P. Stopka/S. Knights (2003), The Puzzle of Human Co-operation, in: *Nature 421*, 911–912

Kagel, J./A. Roth (eds.) (1995), *The Handbook of Experimental Economics*, Princeton

Keeley, L. (1996), *War Before Civilization: The Myth of the Peaceful Savage*, Oxford

Krupp, D./L. DeBruine/P. Barclay (2005), A Cue of Kinship Affects Cooperation in a Tragedy of the Commons. *Human Behavior and Evolution Society Conference Presentation*, Austin

LeBlanc, S./K. E. Register (2003), *Constant Battles: The Myth of the Peaceful, Noble Savage*, New York

Ledyard, J. O. (1995), Public Goods, in: J. H. Kagel/A. E. Roth (eds.), *Handbook of Experimental Economics*, Princeton

Mayr, E. (1961), Cause and Effect in Biology, in: *Science 134*, 1501–1506

Mazur, A./A. Booth (1998), Testosterone and Dominance in Men, in: *Behavioral and Brain Sciences 21*, 353–397

McCabe, K. (2003), A Cognitive Theory of Reciprocal Exchange, in: E. Ostrom/J. Walker (eds.), *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*, New York

— /V. L. Smith (2001), Goodwill Accounting in Economic Exchange, in: G. Gigerenzer/R. Selten (eds.), *Bounded Rationality. The Adaptive Toolbox*, Cambridge

Nowak, M./K. Sigmund (1998a), The Dynamics of Indirect Reciprocity, in: *Journal of Theoretical Biology 194*, 561–574

— / — (1998b), Evolution of Indirect Reciprocity by Image Scoring, in: *Nature 393*, 573–577

Poundstone, W. (1992), *Prisoner's Dilemma: John von Neumann, Game Theory and the Puzzle of the Bomb*, Oxford

Price, M. (2005), Punitive Sentiment Among the Shuar and in Industrialized Societies: Cross-Cultural Similarities, in: *Evolution and Human Behavior 26*, 279–287

— /L. Cosmides/J. Tooby (2002), Punitive Sentiment as an Anti-free Rider Psychological Device, in: *Evolution and Human Behavior 23*, 203–231

— /J. Price/O. Curry (2005), Contribution and Punishment as Reciprocal Altruism in the Public Good Game, in: *Human Behavior and Evolution Society Conference Presentation*, Austin

Rege, M./K. Telle (2004), The Impact of Social Approval and Framing on Cooperation in Public Good Situation, in: *Journal of Public Economics 88(7–8)*, 1625–1644

Rosen, S. P. (2004), *War and Human Nature*, Princeton

Smith, V. L. (2003), Constructivist and Ecological Rationality in Economics, in: *American Economic Review 93(3)*, 465–508

Sosis, R. (2003), Darwin's Cathedral: Evolution, Religion, and the Nature of Society, in: *Evolution and Human Behavior 24*, 137–143

Tajfel, H. (1974), Social Identity and Intergroup Behaviour, in: *Social Science Information 13 (2)*, 65–93

Tinbergen, N. (1963), On Aims and Methods in Ethology, in: *Zeitschrift fur Tierpsychologie 20*, 410–433

— (1968), On War and Peace in Animals and Man. An Ethologist's Approach to the Biology of Aggression, in: *Science 160*, 1411–1418

Tooby, J./L. Cosmides (1989), Evolutionary Psychology and the Generation of Culture: I. Theoretical Considerations, in: *Ethology & Sociobiology 10 (1–3)*, 29–49
— / — (1990), The Past Explains the Present: Emotional Adaptations and the Structure of Ancestral Environments, in: *Ethology & Sociobiology 11*, 375–424
Trivers, R. (1971), The Evolution of Reciprocal Altruism, in: *Quarterly Review of Biology 46(1)*, 35–57
— (1985), *Social Evolution*, Menlo Park
— / — (1994), Reintroducing Group Selection to the Human Behavioral Sciences, in: *Behavioral and Brain Sciences 17 (4)*, 585–654
Wilson, E. O. (1975), *Sociobiology: The New Synthesis*, Cambridge/MA
— (1978), *On Human Nature*, Cambridge/MA
— (1998), *Consilience*, New York
Wrangham, R. W. (1999), The Evolution of Coalitionary Killing, in: *Yearbook of Physical Anthropology 42*, 1–30
Yamagishi, T. (1986), The Provision of a Sanctioning System as a Public Good, in: *Journal of Personality and Social Psychology 51*, 110–6
Zahavi, A. (1975), Mate Selection – A Selection for a Handicap, in: *Journal of Theoretical Biology 53*, 205–14