

The return of reciprocity: a psychological approach to the evolution of cooperation

Alejandro Rosas

Received: 19 October 2006 / Accepted: 19 February 2007 / Published online: 23 March 2007
© Springer Science+Business Media, B.V. 2007

Abstract Recent developments in evolutionary game theory argue the superiority of punishment over reciprocity as accounts of large-scale human cooperation. I introduce a distinction between a behavioral and a psychological perspective on reciprocity and punishment to question this view. I examine a narrow and a wide version of a *psychological* mechanism for reciprocity and conclude that a narrow version is clearly distinguishable from punishment, but inadequate for humans; whereas a wide version is applicable to humans but indistinguishable from punishment. The mechanism for reciprocity in humans emerges as a meta-norm that governs both retaliation and punishment. I make predictions open to empirical investigation to confirm or disconfirm this view.

Keywords Evolutionary theory of cooperation · Experimental economics · Punishment · Reciprocity

Introduction: the limits of reciprocity

Recent developments in the evolutionary theory of cooperation give punishment of free-riders a pride of place in the theory and relegate reciprocity to a marginal role (Fehr and Fischbacher 2004a; Sripada 2005). Since Trivers (1971) proposed reciprocity to explain the adaptiveness of non-selfish behavior towards non-relatives, doubts have accrued concerning both the extent of reciprocity in the biological realm (Noë 1990; Connor 1986, 1995; Clements and Stephens 1995; Stevens et al. 2005), and its ability to explain large-scale human cooperation. Both theoretical models and empirical evidence show that in n -person games with a PD structure where n is large reciprocity causes the decline of cooperation over time, whereas punishment sustains it (Boyd and Richerson 1988, 1992; Fehr and Gächter 2000, 2002).

A. Rosas (✉)
Departamento de Filosofía, Universidad Nacional de Colombia, Ciudad Universitaria, Bogotá,
Colombia
e-mail: arosasl@unal.edu.co

One of the main reasons for distinguishing reciprocity from punishment is the inability of the former to sustain cooperation in the provisioning of public goods in large groups. This fact can be understood intuitively: Suppose a group of 10 people has a norm limiting the amount of water to be drawn daily from a common well. If one of them draws more than the agreed amount, reciprocity causes the other 9 persons to do the same. Reciprocity operates on a rule that prescribes cooperating in the first round of the game, and subsequently doing whatever the other has done in the previous round. This rule succeeds in promoting cooperation in dyadic interactions, but is self-defeating as a mechanism for ensuring cooperation in n -person interactions, because it cannot direct retaliation specifically at defectors. Retaliation in n -person games means defecting not only on defectors, but also on the cooperators present in the group. Punishment is not subject to this deficiency, because it imposes costs on defectors with specific actions directed at them and only at them. So punishment, unlike reciprocity, achieves cooperation in the provisioning of public goods. Empirical evidence suggests that punishment is the mechanism actually used (e.g. Sober and Wilson 1998; Fehr and Gächter 2002; Fehr and Fischbacher 2003; Henrich et al. 2006).

Another reason for the distinction between reciprocity and punishment of free-riders is emphasized by experimental economists. Reciprocity is engaged whenever the reciprocating individual can elicit and predicts future rewards. Altruistic punishment of uncooperative behavior, in contrast, is manifested even if it is costly and provides neither present nor future material rewards for the punisher. Willingness to punish in this sense is widespread in human societies: it has been experimentally confirmed in adults sampled from 15 diverse populations from five continents, including small-scale societies (Henrich et al. 2006).

Reciprocity and punishment are understood in evolutionary theory and in evolutionary game theory (EGT) as strategies in the sense of categories of *behavior*. The former defines them in terms of their functional effects; the latter in terms of utilities or payoff matrices. Though strategies are usually formulated in terms of rules, reference to rules does not pretend to imply intentional mechanisms instantiated in the cooperating individuals. As explained below, strategies abstract from proximate mechanisms in order to allow for a wide application across biological organisms. However, when the proximate mechanisms are *psychological and intentional*, the distinction of strategies at the behavioral level need not match a distinction of mechanisms at the psychological level.

I here raise the issue of the psychological mechanisms supporting the strategies in question, and particularly the question whether the psychology of norm-guided behavior underlies both reciprocity and punishment in humans (Gibbard 1990). Concerning reciprocity, most authors identify it with a *tit for tat*-like strategy. But it is important to realize that this strategy can be driven by simple mechanisms in cognitively unsophisticated creatures, whereas an altogether different mechanism will underlie human reciprocating behavior. Reciprocity in humans is normatively guided and the driving mechanism will include all the features of the psychology of norms, including dispositions to punish and to anticipate punishment. I will therefore argue against matching distinct cooperative strategies in EGT to distinct psychological mechanisms for cooperation in humans.

I will proceed as follows. In the second section I give a narrow construal of reciprocity that allows for a clear distinction to punishment. But such a construal is inadequate for human agents. This will be followed in the third section by an explanation of reciprocity and its driving mechanism as I believe it to be roughly true of humans. The connection between reciprocity and punishment will emerge as inevitable. In the fourth section I illustrate through a simple game some differences between punishment through costly

action and punishment through defection or withholding cooperation. In the fifth and final section, I state some predictions following from the hypothesis that one mechanism underlies both strategies, suggesting their connection to experimental evidence.

The reciprocity-mechanism narrowly construed

I here sketch a version of a rule of reciprocity that fits the claim that reciprocity differs psychologically from punishment. I call it the narrow construal of the rule. The narrow construal makes the distinction between reciprocity and punishment clear cut, but it fails as a plausible version of a mechanism underlying reciprocity in humans.

Before addressing the rule, a clarification is in place. There are two possible perspectives scientists can adopt in describing the behavioral rule followed by a given behavior designed for interaction. One is adopted by evolutionary biologists. When evolutionary biologists formulate the rule for reciprocity saying that the rule requires “cooperation in the first round, thereafter cooperation in response to cooperation, defection in response to defection” they are describing the required behaviors through their evolutionary functions. Neither the idea of a rule, nor the concepts with which it is formulated (cooperation and defection) are meant to capture the nature of the proximate mechanism driving behavior. It is not implied that among the psychological machinery of the “cooperating” organism there is a rule formulated in those terms that is at least a partial cause of its behavior. The standpoint of the evolutionary biologist rests on the judgment that organisms may be capable of exhibiting reciprocal cooperation independently of their level of cognitive sophistication. Differences in their cognitive and motivational machinery leave the similarities in functional behavior unaffected. However, it is essential for reciprocity that the organism be able to respond conditionally to an equivalent behavior in the interaction partner (Axelrod and Hamilton 1981).

From another perspective, the interest lies in determining the proximate mechanisms for “cooperative” behavior. This is the perspective invoked to make the point of this paper. Here it is important to ask what mechanisms drive the conditional responses to an equivalent behavior on the part of the interaction partner. Let us restrict the consideration to organisms that have a nervous system and sense organs. In order to give the mechanism a narrow construal, begin by giving reciprocity a neutral meaning. In a neutral sense, ‘reciprocating’ means responding to an action *A* with an action like it. Organisms interacting in this sense may be said to be reciprocating, even if the actions performed and the interaction itself is not understood by them as cooperative. This lack of “first-person” understanding is characteristic for the narrow construal of the rule.

A narrow construal implies lower cognitive demands. Two organisms reciprocating in the narrow sense only need the ability to carry out behavioral routines perceived to be similar in some respect. These routines must effectively result in either benefit or harm to the other organism, but the organisms themselves need not understand their behavior in terms of reciprocal transfer of benefits. The organisms engage in reciprocating behavior without understanding their actions in terms of cooperation. In order to make their behavior contingent on whether they are harmed or benefited, their reciprocal behavior may be triggered by simple rules appropriate for organisms with nervous and sensory systems. The proximate mechanism could be using rules of this sort: “If *X* (meaning ‘Alter does *x* to you’), then X^R (meaning ‘Do *x* to Alter’).” The rule for reciprocating is not conditional on understanding *X* or X^R as elements in a reciprocal cooperative interaction. Individuals reciprocating in this narrow sense only need to be reinforced in “copying” what the other

has previously done. This is nicely reflected in the second half of a usual formulation of *tit for tat*: “cooperate in the first round, thereafter copy what the other has done.” Nonetheless, the function of the behavior is cooperation, because the ability to elicit benefits by transferring them accounts for its adaptive value.

Empirical evidence suggests that cooperative, *tit for tat*-like interaction in most non-human animals occurs in bouts when the time delay between action and reaction is short (e.g. mutual grooming, egg trading); but when the delay increases, it is rare. This is attributed to the fact that, with short delays, there are no cognitive demands on memory and individual recognition, demands which few non-human animals are able to meet. More importantly, short delays avoid the obstacle of temporal discounting or the devaluing of future benefits. Animals exhibit in contrast to humans high temporal discounting. Experiments with different animal species show that a reward devalues by 50% in 2–6 s (Stevens et al. 2005). This would explain why reciprocation occurs in bouts with short delays between action and reaction. Temporal discounting can be associated to a plausible difference between humans and non-human animals: most non-human animals don’t have the psychological apparatus required to follow prudential rules of action, i.e., rules that have a normative authority derived from a representation of the agent’s overall interests (Fehr and Fischbacher 2004a). Awareness of something like overall interest requires low temporal discounting, i.e., giving significant weight to future benefits.

This narrow construal of the mechanism for reciprocity contrasts with psychological mechanisms working as authoritative conditional rules. For in the narrow sense of “reciprocating”, no norm of reciprocity is being followed, and no norm-violation or “cheating” takes place, except in the metaphorical talk of evolutionary biologists. Trivers famously said, as he introduced reciprocal altruism and its relation to cheating: “Cheating is used throughout this paper solely for convenience to denote the failure to reciprocate; no conscious intent or moral connotation is implied” (Trivers 1971, 36). This usage had to be introduced to allow for cases like the one discussed above, and it applies to ‘reciprocity’ and ‘altruism’ as well. Reciprocity in this narrow sense is non-controversially distinguished from altruistic punishment, but it cannot play an important role in an explanation of large-scale human cooperation.

Reciprocity and punishment: the connection

One way of construing the rule widely is to claim for it what was disclaimed for narrow reciprocity. The concepts used by evolutionary biologists to formulate the rule (cooperation and defection) do capture the nature of the proximate mechanism driving the behaviors of reciprocators widely construed. Humans are an example. Somewhere in their psychological machinery there is a rule formulated in those terms. It is at least a partial cause of their behavior, at least in some people. The concepts of cooperation and defection correspond to representations of a reciprocal transfer of costs and benefits in the proximate mechanisms of cooperators and not only to the evolutionary functions of their behavior. The rule driving behavior is not copying rule, but a *norm* prescribing a given behavior as *good* or *due*. This is crucial in understanding the connection between reciprocity and punishment.

Intuitively, a sharp distinction between reciprocity and punishment is first called into question by noting that the component of retaliation in reciprocal altruism and the punishing action in altruistic punishment have similar effects. One common form of punishment is ostracism or exclusion (Boehm 1999; Bowles and Gintis 2003). The ostracized

individual no longer receives goods or services produced collectively in the group. The costs imposed on the ostracized individual are equal to what he or she loses by exclusion. Similarly, reciprocal altruism prescribes permanently defecting against a defector in a dyadic interaction. This is equivalent to terminating what could have been a cooperative interaction. If the reciprocal altruist can reestablish cooperation elsewhere, whereas this is difficult for the defector due to reputation effects, the defector suffers fitness costs equal to what he or she loses by failing to benefit from cooperation. The effects of both behaviors are similar, despite their different contexts (n -person interaction in the former; dyadic interaction in the latter case). This suggests that the boundary between the intentional mechanisms is thin, if it exists at all. If an abstract rule is designed to impose costs on defectors, the same rule could produce different behaviors with this purpose in different circumstances.

What could this general rule be? Let us assume a theoretically plausible account of norms defended by many social scientists both recently and in the history of the subject (see Sripada and Stich, forthcoming). A psychological state described as “endorsing a norm”, includes an intrinsic motivation to the behavior prescribed by the norm and a disposition to punish norm-violators (and to anticipate punishment). In the wide account, we construe reciprocity as a norm in this sense. Moreover, the norm of reciprocity is plausibly of a special type, according to the views of some social scientists. In “The Norm of Reciprocity”, a paper that counts as one of Trivers’ inspirations for writing his classic piece, a view attributed to Malinowski describes reciprocity as a mechanism governing adherence to norms in general:

Malinowski ... holds that people *owe obligations to each other* and that, therefore, conformity with norms is something they *give to each other* ... (Gouldner 1960; italics in the original)

If we follow this lead, reciprocity emerges as a meta-norm that reveals one important source of the authority of norms, of the grip that norms of cooperation have on us. The meta-norm of reciprocity (hereafter: Reciprocity) demands a conditional compliance with norms: If others comply, we owe compliance to them, because this reinforces and stabilizes compliance and because everyone is better off if compliance is stable. The meta-norm of Reciprocity says something like this:

- (1) Comply in response to compliance; do not comply in response to non-compliance.

Reciprocity, in this sense, is a motivational mechanism that is present wherever we follow norms. Its purpose is to coordinate the long-term interests of different agents repeatedly interacting in a group (ideally, at an efficient equilibrium). But is this mechanism expedient? In virtue of Reciprocity, both cooperation and defection are self-reinforcing. In dyadic interactions, defectors induce reciprocal cooperators to defect. Moreover, and worse, in the case of n -persons interactions for the provision public goods the mechanism seems biased to reinforce defection. According to theoretical models and to public goods experiments, when players defect in absence of punishment opportunities, defection propagates through the group like a wave (Boyd and Richerson 1988; Fehr and Gächter 2000, 2002). Cooperation collapses, and the public good vanishes. The mechanism is incapable of restoring cooperation, but reinforces defection instead.

This is the objection with which we began this paper. It seems a killer, but it is not. It forgets about the human ability to generalize. The claim that reciprocity is unable to sustain cooperation, while direct punishment can, fails to take notice of the fact that cooperation in

public goods interactions can be enforced otherwise than through direct punishment. Specifically, it can be enforced by linking public goods interactions with two-person interactions with the structure of direct or indirect reciprocity games (Milinski et al. 2002; Panchanathan and Boyd 2004). These alternative forms of enforcement raise the question whether a common motivational mechanism underlies all forms of enforcement and whether it is sensible to reserve the term ‘punishment’ to only one form. It is plausible that one and the same motivational mechanism, which I here call Reciprocity, is responsible for generalizing both sensitivity to reputations and retaliation or punishment across the several domains in which agents interact.

The thesis, more fully stated, is this. Since human social life is pervaded by cooperative interactions, victims have several alternative ways of defecting in response to anybody’s defection. Cooperators can thus continue contributing to public goods in the face of free-riders, and yet carry out retaliation against them in other domains. Moreover, retaliation against defectors comes not only from their victims, but also from third parties, who are guided by Reciprocity when retaliating against or punishing defectors that have harmed others (Sugden 1986; Alexander 1987; Fehr and Fischbacher 2004b; Henrich et al. 2006). This is known as indirect reciprocity (Alexander 1987; Nowak and Sigmund 1998), but it follows from Reciprocity, conceived as a generalized meta-norm (Trivers 1971).

Once a defector is marked as such in a group, Reciprocity generalizes retaliation across domains (n -person and dyadic) and engages third parties. In virtue of this double generalization humans succeed in solving the problem of public goods. Somebody who refuses to contribute to public goods will lose his or her reputation in the group. This leads to exclusion from social life. How does exclusion work? Anybody living in a group constantly depends on dyadic interactions in order to cash the benefits of group-membership. Particularly, individual rights depend on general compliance with the corresponding duties. For example, the right to use and dispose of my property depends on others complying individually with their obligation not to seize it at will. If others consistently non-comply, my right is as good as non-existent. Even direct punishment, such as physical harm or even death, implies non-compliance in a dyadic interaction with norms that protect life or personal integrity. Exclusion implies, therefore, denying the norm-violator the recognition of individual rights. In virtue of Reciprocity, defectors on public goods suffer retaliation by cooperators in any other domain of interaction. In this way, Reciprocity is the norm underlying all types of punishment; this finds expression in a well known *dictum* for retributive punishment or justice (“an eye for an eye”).

A simple game for comparing punishment strategies¹

Punishment is normally implemented in dyadic interactions, either in exclusion from cooperation or in direct punishment such as inflicting physical harm. I suggested in the previous section that the same motivational mechanism, namely Reciprocity, drives both retaliation and direct punishment. In this section I explain the difference between them as a difference in the ‘method’ of punishment, a difference motivated by considerations of efficiency. Payoff matrices are assigned to both methods and the payoffs for players are calculated for different scenarios. Differences in efficiency between methods for the different scenarios will emerge. Under the hypothesis that the same motivational mechanism

¹ This section follows advice by an anonymous referee to use pay-off matrices to investigate the differences between reciprocal defection and direct punishment.

underlies both methods, I shall suggest that agents motivated to discipline defectors should also be motivated to switch between methods in varying situations under considerations of efficiency. On this basis, in the following and last section I will make testable predictions about the observable behavior of agents confronted with these decisions.

Assume social life is a complex, indefinitely repeated game composed of alternating rounds of a one-period *n*-person public goods game (PG) and a one-period two-person punishment game. In the first PG round, players are either cooperators or defectors. Cooperators produce a benefit equally distributed among the *n* players at a cost *C* to themselves. Defectors save *C* and benefit as much as cooperators from the PG game. Therefore, cooperators enter round two at a relative negative payoff of $-C$. In the second round agents form pairs and play a one-period two-person game where cooperators punish those who defected in the PG round. To make it simple, cooperators in the PG rounds are identical to punishers in the two-person rounds. Cooperation in the PG round can be enforced if relative payoffs favor cooperation over defection, as is observed in lab experiments (Fehr and Gächter 2000; Milinski et al. 2002).

The two-person game of round two has two versions, each affording a different punishment method. The first method is a two-person punishment game (TPP), where cooperators directly punish those who defected in the PG game at cost to themselves but at an even greater cost to the punished, in a ratio of 1:2. The second punishment method is a two-person interaction with a prisoner’s dilemma structure (TPR). In this TPR game cooperators choose their partners among the other cooperators in the PG game and play cooperatively. Defectors are punished by being excluded from cooperation. The two payoff matrices for the punishment rounds of the game are depicted in Fig. 1a and b.

I consider four different scenarios A, B, C and D for the complex game described above. The first three scenarios vary only the number of players and the proportion of defectors to cooperators:

- (A) 4 players, 2 of them defect in the PG round of the game;
- (B) 30 players, 10 defectors in the PG round of the game;
- (C) 30 players, 20 defectors in the PG round of the game.
- (D) A fourth scenario presents a situation where some defectors pair up with cooperators in the TPR round of the game and are able to exploit them.

Players who cooperate in the PG round enter the punishment round with $-C$; defectors enter this round with 0 (zero). When the second round is a TPR game, cooperators pair up to interact and gain 3 each. Defectors pair with each other and obtain 0 (zero). This applies in scenarios A, B and C, independently of the number of players (provided an even number) and of the proportion of defectors. Cooperation in the PG game can be enforced when

$$3 - C > 0 \tag{1}$$

(a)			
		C	D
C	3, 3	-1, 4	
D	4, -1	0, 0	

(b)			
		not-punish	Punish
not-punish	0, 0	-2, -1	
Punish	-1, -2	-3, -3	

Fig. 1 (a) Payoff Matrix for the TPR game. (b) Payoff Matrix for the TPP game

In the game with the TPP version of punishment we calculate the payoffs separately for each of the three scenarios.

(A) With 4 players and 2 defectors, each defector is punished by the 2 cooperators and obtains -4 ; the two cooperators obtain -2 each. TPP can enforce cooperation in the PG game when

$$2 - C > 0^2 \quad (2A)$$

Comparing (2A) with (1), the TPR structure offers cooperators a greater relative payoff over defectors.

(B) With 30 players in total and 10 defectors, each defector is punished by 20 cooperators and obtains -40 ; the 20 cooperators obtain -10 each. TPP can enforce cooperation in the PG game when

$$30 - C > 0^3 \quad (2B)$$

Relative payoff for cooperators in (2B) suggest that TPP results in over-punishment of defectors. This punishment method would be appropriate if the net cost C of contributing to the PG game were exceedingly high, or in other conditions to be addressed below.

(C) With 30 players in total and 20 defectors, each defector is punished by 10 cooperators and obtains -20 ; the 10 cooperators obtain -20 each. TPP can enforce cooperation in the PG game when

$$-C > 0 \quad (2C)$$

Since C represents the advantage of defectors over cooperators in the PG round and is a positive number, relative payoffs show that TPP cannot enforce cooperation in this scenario. Comparing (2B) and (2C), we infer that TPP cannot enforce cooperation in the PG round when the ratio of defectors to cooperators is above a certain threshold.

These simple scenarios reveal that TPR is a stable and efficient punishment method when defectors cannot exploit cooperators in TPR game. It works independently of the ratio of defectors to cooperators. TPP, in contrast, fluctuates between over- and under-punishment depending on the ratio of defectors to cooperators. Given that in any realistic model of social life TPR interactions among players are always present, it is efficient to punish defectors in PG interactions by excluding them from the benefits of cooperation in TPR interactions, provided it is not possible for defectors to exploit cooperators.

But this last assumption is not always realistic. In real life, it is sometimes impossible to exclude defectors from exploiting cooperators in two-person interactions where benefits can be transferred or withheld. If for any reason, e.g., due to the use of violence or deception, some defectors are able to exploit their interaction partners, then defectors will no longer obtain 0 (zero) in the TPR round. Depending on the opportunities for exploitation, Eq. (1) above will be quickly invalidated and the relative payoffs of cooperators and defectors reversed. In this scenario TPR will no longer be an efficient punishing method.

² If we assume that cooperation is stable if and only if 'payoff to cooperators' > 'payoff to defectors', and then calculate payoffs for cooperators and defectors on the basis of the payoff matrices in Fig. 1a and b, we obtain that cooperation is stable in this scenario if ' $-2 - C > -4$ '. By simple manipulation we arrive at ' $2 - C > 0$ '

³ Again, we arrive at this inequality by manipulating the condition for the stability of cooperation calculated for this scenario: ' $-10 - C > -40$ ' = ' $30 - C > 0$ '.

Predictions and experimental evidence

If one and the same motivation underlies both punishment methods, the difference in efficiency patterns revealed above allows the following general prediction. Agents that are motivated to punish should, if permitted, switch between methods according to the circumstances and under considerations of efficiency.

There is empirical evidence that suggests that people in fact perceive a two-person interaction with a PD structure as a punishment opportunity. This has been shown—indirectly—in an experiment by Milinski et al. (2002). Milinski and collaborators alternated rounds of standard public goods games with rounds of indirect reciprocity games. In these latter rounds each player was potential donor once and a potential receiver once, but never played both roles with the same player. The history of giving in both rounds of the game of each potential receiver was displayed in every indirect reciprocity round for all players. This design created in the players the need to make contributions in the public goods rounds in order to maintain their reputation in the indirect reciprocity rounds of the game. The results show that defectors on public goods are marked with a bad reputation and are punished with defection in dyadic interactions: they lose donations they could have received in the indirect reciprocity game. Interestingly, this is not how Milinski and collaborators interpreted their results. They argued that refusing to give to someone is not the same as punishment and does not need any special punishing rule or motivation (Milinski et al. 2002). In contrast, I contend that this shows that people perceive two-person games with a PD structure as a punishment opportunity. The experiment proves, on the one hand, that the indirect reciprocity game has the same enforcement effect on would-be defectors as direct punishment: it enforces cooperation in public goods situations. On the other hand, those who refused to give in the public goods rounds were more likely to receive nothing in the immediately following rounds of indirect reciprocity, suggesting that participants viewed those rounds as an opportunity to punish their lack of generosity.

One way to test whether this hypothesis is correct is to design experiments where players can switch between direct punishment (TPP) and withholding donations in indirect or direct reciprocity games (TPR), and then test predictions about the behavior of punishers. One prediction concerns their behavior when the circumstances entail the possibility of over-punishment when TPP is the method used. In the model game described in the previous section, the TPP punishment method results in over-punishment in scenarios where many cooperators play with only a few defectors. If in a given experimental setting players are offered the possibility to choose between punishment methods, players should choose TPR over TPP in these cases.

In the context of over-punishment we can make another prediction. If we design an experimental setting where over-punishment would result from TPP as the punishment method, and restrict players in these settings to using this method, we should observe that some players choose to withhold punishment when they judge that defectors have received enough punishment from others. This could be tested with players previously classified as ‘conditional cooperators’, i.e., those that play fair because they cooperate if they experience or believe that others will cooperate as well (Fischbacher et al. 2001; Gächter 2006). This would strongly suggest that their motivation is not to free-ride, but to avoid resource waste in over-punishment.

Predictions can be made also for the context of games where defectors are able to exploit cooperators. In theoretical models this is easily studied by introducing the strategy ALLC; but in experiments with real subjects, this must be done with alternative methods.

The reason is that real subjects playing games where money is at stake fall mostly into two familiar categories: conditional cooperators (roughly 50%) and free-riders (roughly 25%). The behavioral patterns of other subjects are more difficult to classify, but there is no category corresponding to ALLC (Fischbacher et al. 2001; Fischbacher and Gächter 2006; Gächter 2006). In real life defectors can exploit cooperators when reputations are difficult to track. This circumstance could be experimentally simulated. The prediction is that cooperators that anticipate that some cooperators will be exploited in TPR interaction should switch to a TPP method if given the choice, provided the number of defectors is not above a certain threshold. This would provide an explanation why in complex and large societies, where constant mobility and anonymity afford good opportunities for exploiting cooperators, punishment is preferably implemented through direct punishment methods.

These predictions, if tested and confirmed, would provide evidence for the hypothesis that punishment of defectors is motivated by a rule of Reciprocity as explained in section 3. Both when human subjects retaliate against defectors in reciprocity games and when they punish them in direct punishment games, their aim is to discipline defectors by inflicting losses on them. But humans are also sensible to the fact that punishing behavior must be adjusted to varying circumstances in order to impose costs on defectors in an efficient way, i.e., in such a way that defectors will be motivated to cooperative behavior.

One possible objection against this “efficient punishment” view comes from experimental evidence showing that cooperators punish even when they cannot benefit from cooperation provoked in the defector. Third-party punishment provides the clearest evidence for this thesis (Fehr and Fischbacher 2004b). But the efficient punishment view does not contradict this evidence. Punishment, whatever the method used, is efficient when it succeeds in disciplining defectors, whether or not the punisher benefits in the particular case. The evolutionary function of punishing behavior is to prevent defectors to benefit at the expense of cooperators in general. This is achieved when punishment is efficient. ‘Efficient’ does not mean that the punishing individual benefits in the particular occasion of punishment. It means that cooperative behavior in general benefits and achieves stability against defecting behavior.

Punishing is a public good from which all cooperators benefit, regardless of whether they pay the costs. Plausibly, this creates a higher-order free rider problem (Boyd and Richerson 1992; Sober and Wilson 1998), which could eventually drive cooperators who punish to extinction. This problem cannot be discussed here. The point is that punishers driven by Reciprocity do not look primarily at the balance of costs and benefits for themselves. Motivation to punish is unselfish and emotionally driven (Fehr and Gächter 2002; Fehr and Fischbacher 2004b): it targets the breach of a norm and the ungenerous character of the defector (Gintis et al. 2003). This attitude is designed to inculcate norms, to discipline defectors and to make cooperation worth its cost (the cost of renouncing to make suckers out of cooperators) *for cooperators in general*. Benefits to a particular punisher from a particular act of punishment are not required; but they are not, on the other hand, prohibited. When punishers benefit from acts of punishment in repeated games, this does not automatically mean that punishment is selfish. Punishers act *ultimately* to reverse the fitness differentials of cooperators to defectors, and this includes sometimes their own differentials (Price et al. 2002). But this is compatible with the notion that punishers act *proximately* to punish the breach of norms and the lack of generosity. The latter motivation is usually the most reliable way of obtaining the former effect (Frank 1988).

Considerations of efficiency also suggest that when the cost-efficiency relationship of punishing is poor, subjects will consider punishment no longer worth its cost (Fehr and Fischbacher 2004a). We can thus add another prediction: if agents believe that their

punishing action has no chance of making defectors change their ways (whether or not the punisher benefits from the change) they will lose their motivation to punish. Human punishment targets the character and motivation of defectors. As shown in third-party punishment experiments, humans are satisfied when their punishment produces a change in the defector's behavior or motivation, even if they do not expect to benefit directly.

Conclusion

I have shown that the distinction between Reciprocity and punishment cannot be automatically asserted at the level of psychological mechanisms. If similarities in behavioral patterns count over and above dissimilarities in the proximate mechanisms, both humans and some non-human animals instantiate a strategy of reciprocity that is different from punishment. At the psychological level however, only cognitively unsophisticated creatures will require different proximate mechanisms for punishment and reciprocity. In humans, in contrast, both reciprocal behavior and punishment are plausibly expressions of the same psychological mechanism designed to prevent defector strategies to thrive at the expense of cooperators.

Acknowledgements I gratefully acknowledge support by the Philosophy Department and the Research Division of the Universidad Nacional de Colombia, Financial Support Program 2007 for research groups.

References

- Axelrod R, Hamilton WD (1981) The evolution of cooperation. *Science* 211:1390–1396
- Alexander R (1987) *The biology of moral systems*. Aldine de Gruyter, New York
- Boehm C (1999) *Hierarchy in the forest: the evolution of Egalitarian behavior*. Harvard University Press, Cambridge MA
- Bowles S, Gintis H (2003) The evolution of cooperation in heterogeneous populations, SFI working paper # 03-05-031, May 13 (downloaded on the 8th Nov. 2004 from <http://www.santafe.edu/research/publications/workingpapers/03-05-031.pdf>)
- Boyd R, Richerson PJ (1988) The evolution of reciprocity in sizable groups. *J Theor Biol* 132:337–356
- Boyd R, Richerson PJ (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol Sociobiol* 13:171–195
- Connor RC (1986) Pseudo-reciprocity: investing in mutualism. *Anim Behav* 34:1562–1565
- Connor RC (1995) Altruism among non-relatives: alternatives to the Prisoner's Dilemma. *Trends Ecol Evol* 10(2):84–86
- Clements KC, Stephens DW (1995) Testing models of non-kin cooperation: mutualism and the prisoner's dilemma. *Anim Behav* 50:527–535
- Fehr E, Gächter S (2000) Cooperation and punishment in public goods experiments. *Am Econ Rev* 90(4):980–994
- Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* 415:137–140
- Fehr E, Fischbacher U (2003) The Nature of Human Altruism. *Nature* 425:785–791
- Fehr E, Fischbacher U (2004a) Social norms and human cooperation. *Trends Cogn Sci* 8(4):185–190
- Fehr E, Fischbacher U (2004b) Third party sanctions and social norms. *Evol Hum Behav* 25:63–87
- Fischbacher U, Gächter S, Fehr E (2001) Are people conditionally cooperative? Evidence from a public goods experiment. *Econ Lett* 71:397–404
- Fischbacher U, Gächter S (2006) Heterogeneous social preferences and the dynamics of free riding in public goods CeDEX Discussion Paper No. 2006–01 (downloaded from <http://www.nottingham.ac.uk/economics/cedex/papers/2006-01.pdf> on 2/11/06)
- Frank R (1988) *Passions within reason*. The strategic role of the emotions. W. Norton and Co, New York
- Gächter S (2006) Conditional cooperation: behavioral regularities from the lab and the field and their policy implications. CeDEX Discussion Paper 03 (downloaded from <http://www.nottingham.ac.uk/economics/cedex/papers/2006-03.pdf> on 27/10/06)

- Gibbard A (1990) Norms, discussion, and ritual: evolutionary puzzles. *Ethics* 100(4):787–802
- Gintis H, Bowles S, Boyd R, Fehr E (2003) Explaining altruistic behavior in humans. *Evol Hum Behav* (24):153–172
- Gouldner AW (1960) The norm of reciprocity: a preliminary statement. *Am Sociol Rev* 25(2):161–178
- Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, Bolyanatz A, Cardenas JC, Gurven M, Gwako E, Henrich N, Lesorogol C, Marlowe F, Tracer D, Ziker J (2006) Costly punishment across human societies. *Science* 312:1767–1770
- Milinski M, Semmann D, Krambeck H-J (2002) Reputation helps solve the ‘tragedy of the commons’. *Nature* 415:424–426
- Nowak MA, Sigmund K (1998) The dynamics of indirect reciprocity. *J Theor Biol* 194(4):561–574
- Noë R (1990) A Veto game played by baboons: a challenge to the use of the Prisoner’s Dilemma as a paradigm for reciprocity and cooperation. *Anim Behav* 39:78–90
- Panchanathan K, Boyd R (2004) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* 432:499–502
- Price M, Cosmides L, Tooby J (2002) Punitive sentiment as an anti-free rider psychological device. *Evol Hum Behav* 23:203–231
- Sober E, Wilson DS (1998) *Unto others: the evolution and psychology of unselfish behavior*. Harvard Univ. Press, Cambridge
- Sripada CS (2005) Punishment and the strategic structure of moral systems. *Biol Philos* 20(4):767–789
- Sripada CS, Stich S (2006) A framework for the psychology of norms. In: Carruthers P, Laurence S, Stich S (eds) *The Innate Mind: Culture and Cognition*. Oxford University Press, Oxford, pp 280–301
- Stevens JR, Cushman FA, Hauser MD (2005) Evolving the psychological mechanisms for cooperation. *Ann Rev Ecol, Evol Syst* 36:499–518
- Sugden R (1986) *The economics of rights, cooperation and welfare*. Blackwell, Oxford
- Trivers R (1971) The evolution of reciprocal altruism. *Quart Rev Biol* 46(1):35–57