Evolutionary game theory meets social science: Is there a unifying rule for human cooperation?

Alejandro Rosas

Evolutionary game theory meets social science: Is there a unifying rule for human cooperation?

Alejandro Rosas

PII: S0022-5193(10)00090-1
DOI: doi:10.1016/j.jtbi.2010.02.015
Reference: YJTBI5868

To appear in: *Journal of Theoretical Biology*

Received date: 23 September 2009
Revised date: 11 February 2010
Accepted date: 11 February 2010

Cite this article as: Alejandro Rosas, Evolutionary game theory meets social science: Is there a unifying rule for human cooperation?, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.02.015

Evolutionary game theory meets social science: Is there a unifying rule for human cooperation?

Alejandro Rosas[1,2]

Abstract. Evolutionary game theory has shown that human cooperation thrives in different types of social interactions with a PD structure. Models treat the cooperative strategies within the different frameworks as discrete entities and sometimes even as contenders. Whereas strong reciprocity was acclaimed as superior to classic reciprocity for its ability to defeat defectors in public goods games, recent experiments and simulations show that costly punishment fails to promote cooperation in the IR and DR games, where classic reciprocity succeeds. My aim is to show that cooperative strategies across frameworks are governed by a common underlying rule or norm. An analysis of the reputation and action rules that govern some representative cooperative strategies both in models and in economic experiments confirms that different frameworks share a conditional action rule and several reputation rules. The common conditional rule contains an option between costly punishment and withholding benefits that provides alternative enforcement methods against defectors. Depending on the framework, individuals can switch to the appropriate strategy and method of enforcement. The stability of human cooperation looks more promising if one mechanism controls successful strategies across frameworks.[12]

1 INTRODUCTION

For several decades, evolutionary game theory has investigated the evolutionary dynamics of cooperative strategies with computer simulations and analytical methods. Social scientists studying human cooperation with the tools of evolutionary game theory usually understand strategies as standing for such human rules or norms as they encounter in their everyday scientific practice. Taken strictly, talk of strategies and rules in evolutionary game theory makes no reference to the causes of behaviour, neither psychological nor social. Rules stand in the theory for the action patterns displayed towards all possible situations in the game. For example, TIT FOR TAT is the rule: "cooperate in the first move, thereafter copy in move *t+1* whatever your opponent did in move *t*" (Axelrod and Hamilton 1981). However, when it comes specifically to models of human cooperation, rules can be taken to refer to the social and/or psychological causes of behaviour. Human cognition and action involves rules in this causal sense.

The viewpoint of social scientists suggests a question that seems of some consequence for the current debate about the evolution of human cooperation. This paper is devoted to posing and trying to answer it. It is the question whether the cooperative strategies that succeed in different interaction frameworks and are separately investigated in evolutionary game theory could in fact result from one and the same psychological mechanism in a rule-like format, adjustable to the variety of frameworks. Cooperation thrives against defectors in various PD frameworks, namely in two-person repeated games with the *same* person, in two-person repeated games with *different* persons, and in *n*-person (*n*>>2) repeated games with the same or different persons. In each of them cooperative strategies are studied as discrete entities, but perhaps a common rule underlies them all. Adopting a unifying perspective based on the concept of a strategy as a rule, I argue that cooperation across frameworks can be viewed as sharing an underlying common rule prescribing retaliation or punishment against defectors. I also review some experiments that support this same claim. A caveat: though cooperative strategies can in principle occur in any social organism, the claim about a common normative mechanism is raised here only in relation to human cooperation.

2 COOPERATIVE STRATEGIES WITHIN AND ACROSS FRAMEWORKS

Since Trivers' (1971) classic paper and Axelrod and Hamilton's (1981) formal treatment of reciprocity, several cooperative strategies have proven successful at the population level against uncooperative strategies such as ALLD or other selfish strategies in repeated PD games. Cooperative strategies exploit mutual rewards through reciprocity in direct and indirect two-person games; or they cooperate in *n*-person public goods games (PGG) and punish in a linked dyadic game to enhance cooperation in the PGG. Especially in the two-person games, strategies are legion. However, I will introduce common labels for the *successful cooperative* strategies within each type of game. Within two-person repeated games, strategies like TIT FOR TAT, CONTRITE TIT FOR TAT, TIT FOR TWO TATS and others that exploit reciprocity against ALLD or against conditionally uncooperative strategies, are here labelled direct reciprocity (DR) strategies. Their common element is to cooperate in repeated interactions with the same player, when the player cooperates and avoids cheating behaviour. Analogously, cooperative strategies that succeed through reciprocity in the evolutionary dynamics of repeated two-person games where players never meet twice in exchanged roles are labelled indirect reciprocity (IR) strategies. They cooperate with opponents if these have a good reputation; otherwise they defect. IR strategies are sometimes called after the reputation rule they follow, e.g., SCORING, STANDING and JUDGING; more recently, some theorists refer to reputation rules as social norms and reserve 'strategy' for the action rule that prescribes an action

[1] KLI for Evolution and Cognition Research, Altenberg, Austria. Email: **alejandro.rosas@kli.ac.at**

[2] Dept. of Philosophy, Univ. Nacional de Colombia, Bogotá, Colombia. Email: **arosasl@unal.edu.co**

for each possible game situation. Proponents of strong reciprocity (SR) regard it as a single strategy. The various possible combinations of reputation rules with punishment have not been an issue for theorists until very recently (see Ohtsuki et al. 2009).

The existence of multiple strategies within the first two frameworks may seem, at first sight, to contradict the aim of this paper. The common labels may seem an arbitrary device to artificially create unity where there is none. However, this paper is not concerned with reducing multiplicity of cooperative strategies *within* game-types; it is concerned with reducing it *across* game-types. Unity across game-types is compatible with variation within them. Suppose that there are two irreducible cooperative strategies within the direct reciprocity framework. As long as these two strategies have analogues within the indirect and the strong reciprocity frameworks, rules will be common across frameworks.

The merits of the cooperative strategies across frameworks are often discussed in a competitive attitude: for example, when direct or indirect reciprocity are compared to strong reciprocity and the superiority of the latter is defended or contested (Gintis et al. 2003; Burnham and Johnson 2005; Ohtsuki et al. 2009). Typically, proponents of SR have claimed its unique success against free riders in public goods games. Though PGGs remain an important justification for the existence of SR strategies, this claim needs revision in the face of some experiments (Milinski et al. 2002; see 8 below). The claim that SR stands out as a biologically and psychologically altruistic strategy (Gintis et al. 2003) against the selfish character of IR and DR strategies often accentuates the rivalry. But the altruistic character of SR has been justly questioned (Haley and Fessler 2004; Quervain 2004; Rosas 2008a). Moreover, discovering a common rule to all cooperative strategies across the different frameworks would surely deflate the rivalry. These strategies share a common concern with deterring free riders, which is, e.g., irrelevant in cooperation through by-product mutualism or kin selection.

3 STRATEGIES AS RULES OF TWO TYPES

Taken cognitively, cooperative action rules establish in propositional format whether to give or not, depending on the reputation of the recipient and/or the actor. Taking both reputations into account, there are four possible game situations to consider and a rule consists in a string of four values that establish whether to cooperate or defect in each of these four situations. There are 16 (2^4) such strings or action rules. Not all strategies take both reputations into account. CO is the typical cooperative action rule for IR that looks only at the recipient's reputation and prescribes to give to the good, and not to give to the bad. But if the actor's reputation depends on its action, it is in the actor's interest to look at its own reputation before acting (Leimar and Hammerstein 2001; Brandt and Sigmund 2004). SELF is the uncooperative action rule that dictates cooperation only when the actor's reputation is bad. Action rules can also take account of both the actor's and the recipient's score. AND is the rule that prescribes to give only when the recipient is good and the actor is bad. OR prescribes to give when either the recipient is good or the actor is bad. Simulations in Brandt and Sigmund (2004) and analyses in Ohtsuki and Iwasa (2004) show that SELF and AND are not evolutionary successful, whereas CO and OR achieve high fitness. CO and OR may be irreducible cooperative rules; yet our goal here is not to eliminate variation within a game's framework, but to find out if there are common rules shared by cooperative strategies across frameworks.

A conditional action rule like CO will operate together with a reputation module, which assesses the reputation of potential recipients according to a rule. In most models reputations are simplified to binary scores– either good (G) or bad (B) – though in real life reputation is assigned in a continuous scale. For the discussion here I will assume the simplified binary model. The rule must take as input the previous action of the potential recipient, and possibly the reputation of its respective recipient and/or its reputation when it moved as donor. Taking all three factors into account there are eight possible actions to assess. A reputation rule is a string of eight values giving the reputations, G or B, for these eight actions. In total there are 256 (2^8) "third order assessment" rules. Multiplying 256 times the 16 possible actions rules results in 4096 possible strategies just for the IR game type (Ohtsuki and Iwasa 2004). Fortunately for our purpose, the number of relevant strategies is much smaller. Not all action rules are cooperative, and not all possible reputation rules promote cooperative action rules. For example, the cooperative action rule known as CO, which prescribes action on the basis of the reputation of the recipient only ("give to the good and do not give to the bad") will undermine itself if it uses SCORING, a reputation rule that assigns scores only on the basis of past actions. SCORING contradicts CO because it assesses a potential recipient as bad if it previously refused to give to the bad, and as good if it gave to the bad.

Because of this problem Ohtsuki and Iwasa (2004) carried out an exhaustive search of evolutionary stable combinations of reputation rules with cooperative action rules. They found eight such ESS combinations that exclusively use cooperative action rules CO (six) and OR (two) (under OR, a bad actor will cooperate with a bad recipient to become good) and reputation rules like STANDING and JUDGING. In simulations of the dynamics and the ability to invade populations of defectors, Brandt and Sigmund (2004) found that three reputation rules suggested by common moral intuition (SCORING, STANDING and JUDGING), support the success of cooperative action rules CO and OR. Their success was twice as high with JUDGING than with SCORING. Later, Ohtsuki and Iwasa (2007) investigated the dynamics of three action rules: ALLC, ALLD and CO (the latter labelled DISC in their paper) under 16 second-order reputation rules and found two reputation rules that promote CO: STERN JUDGING (KANDORI in their paper) and SIMPLE STANDING. Since ALLD and ALLC have no use for a reputation module, their study investigates the ability of the combination of CO with these reputation rules to invade ALLD and ALLC. A previous study by Pacheco et al. (2006) had investigated the dynamics of all 16 possible action rules under individual and group selection and found that the same reputation rules as in Ohtsuki and Iwasa (2007) promote cooperation. In a different study that investigates the emergence of cooperative strategies out of non-cooperative ones through group selection and probabilistic rather than deterministic rules, Scheuring (2009) found that the dominant cooperative norm is GENEROUS JUDGING, a norm where defection against the bad is rewarded less frequently than in STERN JUDGING, and cooperation with the bad is as likely to be punished as to be rewarded.

Cooperation was promoted in these studies mostly in the form of CO. CO is the conditionally cooperative strategy that can be stated in English as "Cooperate with the good, defect against the bad, independently of your own reputation." The reputation rules that promote CO are second-order rules because they only need to take into account the action of the actor and the reputation of the recipient. STERN JUDGING fits exactly to CO, for it assigns a good reputation to those and only to those that follow CO. In English, STERN JUDGING says: "Cooperate with the good and defect against the bad and you will be good; otherwise you will be bad".

Given these convergent results and our interest in unity of rules across frameworks, I shall focus on CO and ask whether it is shared across frameworks. CO takes account of the recipient's reputation only. There are 4 possible actions of this sort: C-G; D-B; C-B; D-G (cooperate with the good; defect against the bad; etc). Four strings, each combining two of these four actions, constitute the possible action rules. The only cooperative rule that can beat defectors is CO: the string C-G, D-B. Further, I shall ask whether three second-order reputation rules are shared. Table 1 gives the three reputation rules and their assessment of actions and rules. ALLC and ALLD are compared to CO as part of its usual environment of competition.

RecipientÕs reputation	*Action rules*			*Assessment of actions to the left, according to the reputation rule to the right*			*2nd order reputation rules*	*Comments*
	ALLC	ALLD	CO	ALLC	ALLD	CO		Differ in their assessment of actions to the bad (2nd row)
Good	C	D	C	G	B	G	SCORING	Rewards C-B, punishes D-B
Bad	C	D	D	G	B	B		
				G	B	G	SIMPLE STANDING	Rewards C-B, rewards D-B
				G	G	G		
				G	B	G	STERN JUDGING	Punishes C-B, rewards D-B
				B	G	G		

Table 1. Three action rules and their assessment by three second-order reputation rules.

In table 1, CO is the only action rule that cooperates conditionally on the reputation of the recipient. The middle column shows the assessment of the actions and action rules on the left according to three second-order reputation rules listed on the right. These rules only take into account the action of the actor and the reputation of the recipient, except SCORING that only cares for the action. The farthest right column describes the differences between the reputation rules, highlighted in grey. SIMPLE STANDING is the most generous reputation rule because it totally approves of ALLC and CO; STERN JUDGING is more generous than SCORING and, crucially, more consistent in the promotion of conditional cooperation.

4 ACTION RULES

The question then is whether the cooperative strategies belonging to direct, indirect and strong reciprocity frameworks are constructed with CO and the reputation rules mentioned above.

CO underlies most DR strategies, namely, those that prescribe giving to good opponents and not giving to bad ones, e.g., TFT and Contrite TFT. As we shall see, these differ in the criteria that mark an opponent as bad (the reputation rules), but the same difference appears within IR as well, namely between SCORING and SIMPLE STANDING. If DR and IR cooperative strategies are both based on the action rule CO, and if they share similar reputation rules, how do DR and IR strategies differ? DR strategies apply only when the interaction between the same two players is repeated; IR strategies apply only when repeated dyadic encounters are excluded. In this way, no overlaps occur between DR and IR strategies and evolutionary game theory can assess their evolutionary dynamics separately.

In fact, this is the only difference between DR and IR cooperative strategies; and the question arises whether the underlying rules should be different only for this reason, in particular since pure IR interaction domains are suspiciously artificial. Traffic in the streets may adequately reproduce the typical IR one-shot structure; but it is hopeless to try to keep updated about the reputations of traffic participants. In domains with a smaller circle of participants you cannot exclude repeated interaction. In this case, simulations show that IR strategies are not always stable against DR strategies. DR/TFT beats IR/SCORING (see 6 and 7 below) when pairs meet repeatedly (40 times) both in small (10) and in large (100) groups (Roberts 2008). The reason is that IR/SCORING punishes co-operators that justifiably defect on defectors, whereas DR identifies co-operators better through direct interaction. In contrast, IR/STANDING avoids punishing punishers (see sect. 6) and beats DR/CTFT (contrite TFT, see sect. 7) in the same environment. Interestingly, though it may seem that IR and DR strategies are competing in those simulations, the alleged IR players in that model are in fact switching between IR and DR as required. Switching seems unavoidable in natural contexts and raises the question whether common rules facilitate it.

In Roberts' simulations players interact in pairs and any two of them, *A* and *B*, will meet about 40 times. In this model any player's memory extends only to one period. DR players remember only the opponents' last move towards them. But IR players remember the last move, no matter whether it was towards them or towards others. Thus, if *A* is an IR player, it will switch in this model between DR and IR, depending on the target of *B*'s last move. Switching is easier if the player applies the same action and reputation rules in both frameworks, but it is still possible under different rules if the inputs univocally trigger their corresponding rules. However, if we add long-term memory, the chances that *A*'s memories of any given agent's actions are either only towards *A* or only towards others will be very low. If the actions and the memories are mixed, rules for DR and IR will be triggered and, if different, a potential conflict will ensue. Consistency is maintained if the same reputation rules and the same action rules are shared across DR and IR contexts. Sharing rules means that reputation rules can process inputs (memories) of both types and output a reputation that is treated by a common action rule indistinctly of origin. This sharing of rules across frameworks must explain in part the evolutionary scope and stability of human cooperation.

5 A GENERAL ACTION RULE

I shall now formulate a general rule that also encompasses SR strategies. Cooperative rules need to cope with defectors if they are to succeed (Trivers 1971). The action rule CO in DR and IR prescribes to withhold benefits from them, and succeeds at the

population level in dyadic games. But if defectors cannot be specifically targeted with defection, as happens in public goods games, investment in costly punishment (SR) is a viable option. To punish is to respond to a defection with spite in a separate dyadic interaction, where the cost to the punisher is usually less than the cost inflicted on the punished (Fehr and Gächter 2002). This form of punishment is more severe than withholding benefits, but both costly punishment and withholding benefits legitimately count as forms of punishment. To avoid confusions, I will use 'enforcement' as the umbrella concept to cover both withholding benefits and costly punishment.

When the structure of the game does not allow targeting defectors specifically, a dyadic interaction must be added for this purpose (Boyd and Richerson 1992). When SR is invoked in connection to public goods, ultimatum and third-party punishment games (Fehr and Fischbacher 2004a,b) the underlying structure consists in two games linked in two stages. The SR strategy comes into play as spite in a dyadic second stage, directed to a player perceived as violating a norm of cooperation in the first stage. Strictly speaking, the domain of SR is a dyadic interaction linked as second stage to a social dilemma interaction – either dyadic or *n*-person, *n>>2* – in a first stage. This two-stage structure can be repeated. Repeated PGGs with punishment are sometimes played in the lab reshuffling the groups after each period, so as to imitate one-shot encounters (Fehr and Gächter 2000). But this is instrumental to proving that punishment is biologically altruistic. In real life, SR strategies are more often played in stable groups. I here define SR, contrary to the will of its initial proponents, as involving costly punishment, but not necessarily altruistic punishment, because the punisher may receive compensation either through the future cooperation of the punished or from third parties interested in promoting cooperation. The only difference between SR strategies and DR/IR strategies concerns the part of the cooperative action rule that instructs how to enforce against defectors. The common action rule instructs to cooperate with co-operators and to enforce against defectors, allowing a choice between spite (SR) or withholding cooperation (DR/IR). The question arises, naturally, why these two methods of enforcement and whether both are really needed. The answer is that they are. Costly punishment should be engaged only when withholding benefits is unable to enforce cooperation; but then it *must* be engaged. Experiments suggest that this happens more often in connection with public goods (Rockenbach and Milinksi 2006).

6 REPUTATION RULES

Reputation rules are often discussed as if they were exclusive to IR. But if they are the criteria that mark players as good or bad conditional on their behaviour – rules taking behaviour as input and delivering a reputation as output – all conditional strategies must use them. Whether the behavioural data come from direct interaction as in DR and in public goods games, or from observation and hearsay as in IR, the rules for processing input data can be the same. Here, again, variety within frameworks is to be distinguished from variety across frameworks. If reputations rules are diverse, but the same diversity is relevant in all frameworks, it is likely that any individual using a particular rule within a framework will see reason to use the same rule across frameworks.

Reputation rules have been developed extensively in relation to IR (Brandt and Sigmund 2004; Ohtsuki and Iwasa 2004; 2006; 2007). An actor *A* meets a potential recipient *B*. *A* first evokes a memory of *B*'s last move as actor towards a third party *C*. Reputation is then allocated as a binary score, i.e., good or bad, according to a rule. The simplest rule considers only whether *B* helped *C* or not; this rule is called SCORING. Other rules take into account the reputation of the recipient *C* (second-order) and yet others take also into account the reputation of the observed actor *B* (third-order). SIMPLE STANDING and STERN JUDGING are two second-order rules. They assign a good reputation to *B* for refusing to help *C* if *C* was bad. STERN JUDGING additionally gives a bad score to cooperation with a bad recipient. In the former case, defection against a bad recipient is viewed as justified defection or punishment. In the latter, cooperation with a bad recipient is seen as the omission of due punishment, i.e., as treason.

7 REPUTATION RULES ACROSS FRAMEWORKS

Important here is that these rules are applied across the different frameworks. Reputation rules for DR can be seen as special instances of rules for IR: when *A* evokes a memory of a past interaction of *B* with *C* such that *C=A*. TFT uses SCORING: it only considers whether *B* defected or cooperated. With this simple rule, two TFT players can lock into mutual punishment in a sequential two-person game. This happens when *A* makes a mistake and defects, provoking *B*'s defection, which *A* does not recognize as justified. This motivated CTFT (contrite tit for tat), a strategy that recognizes its errors of implementation, so that in the previous case *A* returns to cooperation (Sugden 1986; Boerlijst et al. 1997). CTFT applies the same rule as SIMPLE STANDING: *A* will not mark *B* as having a bad reputation for refusing to help *A* if at that moment *A* was bad. *A* interprets *B*'s defection as justified punishment. It may seem harder to make sense of equivalents for STERN JUDGING in DR, which would mark an opponent as bad for cooperating with *A* when *A* had defected in the previous round. This form of "insistent cooperation" matches practices that we know as flattery or obsequiousness. Though this behaviour is sometimes accepted by the flattered, it is sometimes rejected. Sometimes, the targets of flattery or obsequiousness react with contempt and anger because those behaviours betray either weakness or the intention to manipulate. In this case their reputation rule assesses such "co-operators" as bad.

SIMPLE STANDING and STERN JUDGING have hardly played any role in the modelling of SR in repeated public good games with punishment. A punisher in the second stage of these games only needs SCORING to assess reputations of players in the first stage, because defecting in the PGG stage is a sure sign of being a defector. If a player wants to punish defectors, it should punish them in the second two-person stage, not by defecting in the first *n*-person stage. Similarly, cooperating in the PGG stage cannot be interpreted as treason, for treason is omitting punishment of defectors in the second punishment stage. In one of the first simulations of altruistic punishment, Axelrod (1986) modelled the punishment of non-punishers in a third stage, which treats the omission of punishment in the second stage as treason and uses the same reputation rule as STERN JUDGING. Another way to introduce these second-

order rules in SR would be to model an IR game where players have the option of costly punishing defectors (SR) instead of withholding benefits from them. In this case, SR players would not want to punish punishers (STANDING); and perhaps they would want to punish traitors (STERN JUDGING).

This last possibility has been explored recently in some experiments and simulations, where punishment is stabilized by norms that dictate cooperation with those who punish the bad (SIMPLE STANDING) and punishment for those who either cooperate with or withhold benefits (defect) from the bad (similar to STERN JUDGING). The overall conclusion of these studies, however, is that costly punishment does not promote cooperation in DR/IR environments better than withholding benefits (Ohtsuki et al. 2009; Dreber et al. 2008; Rand et al. 2009). This is not surprising, for it is intuitively not advisable to use costly punishment where withholding benefits is otherwise sufficient to enforce cooperation. The use of costly punishment must be justified by peculiarities of the interaction that entail that withholding benefits is not open or not efficient to achieve enforcement. Public goods games are such cases. Surely, withholding cooperation has a similar effect to costly punishment when pair-wise DR or IR interactions are linked to PGGs (Milinski et al. 2002; Barclay 2004). But this arrangement is not always feasible in real life: if defectors can intimidate others into transferring benefits, the few that can resist intimidation will not achieve deterrence without costly punishment. Also, contribution to public goods in modern societies (paying taxes) is not publicly observable. Only institutional agents using costly punishment are here able to enforce. These special circumstances have not been taken into account in those recent studies: they have arranged a competition between withholding benefits and costly punishment in DR and IR environments without any special circumstance that would justify the introduction of costly punishment. Costly punishment is appropriate when withholding benefits is not effective; and when it is appropriate, it is important to remember that the same reputation rules that promote cooperation in IR games will promote it too when costly punishment is advisable.

SCORING, SIMPLE STANDING and STERN JUDGING do not reflect any difference between the frameworks. They reflect, rather, differences in moral assessments that are relevant within any framework (see table 2). In conclusion, the fundamental challenge for unifying strategies across frameworks is the distinction between DR/IR on one hand, and SR on the other. I have suggested that this distinction rests on a common ground: a general rule that prescribes enforcing cooperation with two enforcement options.

Data for reputation rule	*Reputation assigned to B by three reputation rules (St and J are 2nd order rules)*			*Action rule CO across frameworks, dictates A's move towards B given B's reputation*		
B's move to *C* (*A*) | Reputation of *C* (*A*)	Scoring (Sc)	Simple Standing (St)	Stern Judging (J)	DR (Sc, St, J)	IR (Sc, St, J)	SR (Sc, St, J)
C | Good	G	G	G	C, C, C	C, C, C	C, C, C
C | Bad	G	G	B	C, C, D	C, C, D	C, C, P
D | Good	B	B	B	D, D, D	D, D, D	P, P, P
D | Bad	B	G	G	D, C, C	D, C, C	P, C, C

Table 2. Reputation and action rules across frameworks

In table 2, actor *A* assigns a reputation (G=Good; B=Bad) to a recipient *B* according to three possible reputation rules based on two sorts of data: 1) *B*'s last move (C or D) in whichever game *B* was previously involved, e.g., towards a recipient *C*, or towards *A* itself when *A* was *B*'s recipient in an iterated two-person PD, or towards both in an *n*-person PD; and 2) the reputation of *C* or *A*. SCORING only registers *B*'s move; SIMPLE STANDING and STERN JUDGING also pay attention to the reputation of *C* or *A* as *B* moves. The action rule tells *A* to give (C) when *B* is good and to withhold a benefit (D) or costly punish (P) when *B* is bad. Differences between reputation rules concern the assessment of actions to the bad, and they affect all frameworks in the same way (highlighted in grey). Only the action rule determines a difference for frameworks between enforcing through P or through D. The circumstance dictates whether P is justified to achieve effective deterrence.

8 ENFORCEMENT AND REPUTATION IN EXPERIMENTAL GAMES

In this section I review economic experiments that suggest that a common normative mechanism underlies cooperative strategies across the three frameworks under consideration.

SR was introduced upon the discovery, both experimental and theoretical, that in the *n*-person PD with *n >> 2* defecting in response to defectors causes the breakdown of cooperation, whereas responding spitefully in a linked game increases cooperation levels (Kim & Walker 1984; Yamagishi 1986; Boyd and Richerson 1988, 1992). It rose to popularity when experimental economists specifically excluded in the lab – in anonymous one-shot interactions – all benefits that could have been caused to punishers through inducing the punished to cooperate in subsequent periods. Results showed that subjects still punished under these conditions, when no benefit could be expected, and achieved high levels of cooperation (Fehr and Gächter 2000; 2002). Given the importance of public goods for human large-scale cooperation, the success of 'altruistic' punishment arose enthusiasm. But this success was soon rivalled by the performance of an IR strategy in a similar game.

An IR strategy succeeded in promoting the provision of public goods in an experiment by Milinski et al. (2002). In their design, withholding benefits in dyadic, one-shot encounters figured instead of costly punishment in a second stage after the PGG. The experiment revealed that the superiority claimed for SR rested partially on a misunderstanding. Supporters of SR said that DR was expressed in an iterated PGG when co-operators responded with defection to defection, which caused cooperation to unravel. But, just as enforcement of cooperation through costly punishment requires the addition of dyadic game in a second stage after the PGG, a fair comparison with DR should keep the same design. In the added second stage, co-operators can punish defectors dyadically - the game-theoretic context for a DR strategy - by withholding benefits from them or by not choosing them for interaction at all (Rosas 2008b). Milinski et al. (2002) experimented with an IR game in the second stage; an experiment with DR in the second stage has also been conducted, with the same effect (Barclay 2004). Panchanathan and Boyd (2004) produced the model corresponding to Milinski's experiment and found that IR can enforce cooperation "without the second-order free rider problem", i.e., without biological altruism. The fundamental insight, arguably, is that

players can either withhold benefits or costly punish to enforce cooperation in the first-stage PGG, provided both methods appear in the second dyadic stage of a two-stage design.

The moral to this story is that cooperative strategies share common grounds across the DR, IR and SR frameworks. A cooperative strategy cannot survive without retaliating against defectors (Trivers 1971). Across the frameworks, all strategies do this, differing only in their method. This justifies introducing the term 'norm' for strategies in IR frameworks (Nowak and Sigmund 2005), though advocates of SR sometimes write as if norms were exclusive to SR (Fehr and Fischbacher 2003, Gintis et al. 2003). Moreover, in third-party punishment experiments third parties punish defectors in an observed one-shot two-person PD. Advocates of strong reciprocity interpret this behaviour as enforcing a norm (Fehr and Fischbacher 2004a). So, obviously, these third parties view the two-person PD as governed by a norm and the defectors as norm-violators. There is no reason to believe that the participants in the two-person PD view their interaction differently. Cooperation in all social dilemmas is norm driven, not only in cases where SR is required. However, the enforcing power of DR and IR should not suggest that they could in every case replace SR. It remains true that in some cases it is necessary to costly punish to succeed in enforcing cooperation.

An experiment by Rockenbach and Milinski (2006) shows that people understand the enforcing power of both IR and SR strategies. It also suggests that they do not readily renounce to costly punishment in the context of public good provisioning. The experiment studies, in two treatments, enforcement of contribution to public goods. Subjects can either contribute all or contribute nothing. One treatment has two stages: subjects participate in a public goods game (first stage) having chosen beforehand whether punishment will be possible in a second stage. The other treatment has three stages. The first two stages are as just described. Then both groups (with and without punishment) play an IR game where each participant is once an actor and once a recipient. In this way, the experiment obtains data for three different methods to discipline those who defected in the public goods game: 1. combining SR with IR; 2. only through IR; 3. only through SR. Results show that subjects confronted with the choice between 1 and 2 prefer 1. Combining SR with IR enforces a higher number of full contributions than either SR or IR alone and it happens at a lower cost in punishment points than in SR alone. Thus, the combination of SR with IR boosts the efficiency of cooperation for the whole group. This combination builds "a low cost solution to social dilemmas" (ibid.), providing an apt target for natural selection. The experiment suggests not only that SR and IR express and enforce the same norm for conditional cooperation, but that in connection with public goods games people feel the need to apply costly punishment. The norm: "Cooperate with others in your group; enforce against defectors", leaves open whether enforcement should happen as costly punishment or as withholding benefits. But, in connection to public goods games, a combination of both is apparently best to enforce the maximum contribution at the lowest possible cost.

Experiments provide general evidence to confirm that also reputation rules are shared. When subjects play public goods games linked to two-person games, where they can either act on SR or IR or both (Rockenbach and Milinski 2006); or when third parties observe a two-person DR game and punish defectors (Fehr and Fischbacher 2004a), they in fact treat reputations as shared by strategies across different frameworks. Players acquire bad reputations in an *n*-person PGG or in a two-person DR game; and then this reputation triggers responses in subjects playing either SR or IR strategies. Subjects use the same criteria for reputation across game types and across strategies. SCORING, STANDING and JUDGING reflect differences in moral assessment; but the differences apply to any game type and a player will probably have reason to stick to one criterion when moving across game-types (see table 2).

In models, SIMPLE STANDING and STERN JUDGING promote cooperation better than SCORING. But an experiment by Milinski et al. (2001) showed that subjects use SCORING even when second-order information suitable for the use of STANDING was available. A possible explanation is that the experiment hindered the use of STANDING by the way the information about the actions was presented. The information did not come nor was easily transformable into the format C-G; D-B (cooperate with the good; defect against the bad; etc), which is the format of the input for second-order assessment strategies. The information that gives the reputation (G or B) of the receiver of the potential receiver's action being assessed was not given. It had to be inferred from histories of cooperative or defective actions taken from a network created as seven players interacted randomly and continuously in pairs for 16 rounds. This represented an impossible task for working memory, as the authors acknowledge. The question is whether this impossible task represents what is normally required in real life to assess reputations. The authors seem to believe that it does, but I would cautiously look forward to further experimentation with input information presented in a more appropriate format.

9 CONCLUSIONS

Analyses of strategies as rules in both models and economic experiments suggest that a common rule prescribing cooperation with the good and enforcement against the bad underlies all strategies. The common rule takes information about individual reputations as input and outputs decisions either to cooperate or to enforce. This mechanism is restricted to norms of cooperation and is less abstract than the universal normative mechanism advocated in Sripada and Stich (2006). Yet it is sufficiently abstract to encompass all frameworks of interaction (see fig. 1). Theoretically, this hypothesis could be modelled by constructing a type of agent that withholds cooperation from defectors and resorts to costly punishment only when it is not possible to enforce through withholding benefits, for example, in public goods games that cannot be linked naturally to DR or IR games in a second stage. Punishment would also be necessary if defectors were capable of intimidating most players and purloining benefits in two-person games. The few agents that could resist intimidation would have to inflict costly punishment on the defectors; withholding benefits would not effectively enforce. The fact that the norm relies heavily on the ability to see the similarities across frameworks probably restricts its domain to humans. An important consequence for the evolutionary game theory of cooperation is that the question of stability can be posed in relation to the common norm rather than to each strategy individually. The stability of human cooperation looks more promising from this perspective than from the viewpoint of any one strategy considered in isolation.

***A*'s action module (CO)**		Cooperate with the good, enforce against the bad					
***A*'s Reputation module (STERN J)**		Assigns reputation to a potential recipient *B* depending on its past actions					
Input Reputation module: *B*'s past actions		C-G; D-B			D-G; C-B		
Input action module	**Output reputation module: B's reputation**	G			B		
	Interaction *A*-*B*	DR	IR	SR	DR	IR	SR
	Efficient enforcement	No enforcement needed			D	D	P
Output action module: Appropriate action		C	C	C	D	D	P

Figure 1. The basic cognitive architecture underlying cooperative strategies.

A module for assigning reputations (here STERN JUDGING) embedded in a cooperative action module is fed with beliefs about *B*'s past actions and outputs a reputation for *B*. The action module is fed with beliefs about reputation, interaction types and efficient enforcement and outputs the appropriate action (last row). G = good; B = bad; D = withhold benefits; P = costly punishment; C = cooperate. DR = direct reciprocity; IR = indirect reciprocity; SR = strong reciprocity.

Acknowledgments: I thank the KLI for Evolution and Cognition Research and the Universidad Nacional de Colombia for funding. For helpful comments on previous drafts I thank Karthik Panchanathan, Francesco Guala and two anonymous referees for the JTB.

REFERENCES

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

[2] Axelrod, R. 1986. An evolutionary approach to norms. The American Political Science Review 80 (4):1095-1111.

[3] Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons". Evolution and Human Behaviour 25: 209-220.

[4] Boerlijst M C, Nowak M A, Sigmund K. 1997. The Logic of Contrition. Journal of Theoretical Biology 185: 281-293.

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

[6] Boyd R, Richerson P J. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethology and Sociobiology 13: 171-195

[7] Boyd R, Gintis H, Bowles S, Richerson P J. 2003. The evolution of altruistic punishment. PNAS 100(6): 3531–3535

[8] Brandt H, Sigmund K. 2004. The logic of reprobation: assessment and action rules for indirect reciprocation, Journal of Theoretical Biology 231: 475–486

[9] Burnham T C, Johnson D P. 2005. The Biological and Evolutionary Logic of Human Cooperation. Analyse & Kritik 27: 113-135

[10] Dreber A, Rand D G, Fudenberg D, Nowak M A. 2009. Winners don't punish. Nature 452: 348-351

[11] Fehr E, Gächter S. 2000. Cooperation and punishment in public goods experiments. American Economic Review 90(4): 980-994.

[12] Fehr E, Gächter S. 2002. Altruistic punishment in humans. Nature 415: 137–140.

[13] Fehr E, Fischbacher U. 2003. The Nature of Human Altruism. Nature 425: 785-791.

[14] Fehr E, Fischbacher U. 2004a. Third party sanctions and social norms. Evolution and Human Behavior 25: 63-87.

[15] Fehr E, Fischbacher U. 2004b. Social norms and human cooperation. Trends in Cognitive Sciences 8(4): 185-190.

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

[17] Haley K J, Fessler D M T. 2005. Nobody's watching? Subtle cues affect generosity in an anonymous economic game. Evolution and Human Behavior 26: 245–256

[18] Kim O, Walker, M. 1984. The free rider problem: experimental evidence. Public Choice 43:3-24.Kollock P. 1993. An eye for an eye leaves everyone blind: Cooperation and accounting systems. American Sociological Review 58(6): 768-786.

[19] Leimar O, Hammerstein P. 2001. Evolution of cooperation through indirect reciprocation. Proc. R. Soc. Lond. B 268: 745–753.

[20] Milinski M, Semmann D, Bakker TCM, Krambeck H-J. 2001. Cooperation though Indirect Reciprocity: Image Scoring or Standing Strategy, Proc. R. Soc. London B. 268: 2495-2501

[21] Milinski M, Semmann D, Krambeck H-J. 2002. Reputation helps solve the 'tragedy of the commons'. Nature 415: 424-426.

[22] Nowak M A, Sigmund K. 1998. The dynamics of indirect reciprocity. Journal of Theoretical Biology 194(4): 561–574.

[23] Nowak M A, Sigmund K. 2005. The evolution of indirect reciprocity. Nature 437:1291-1298.

[24] Ohtsuki H, Iwasa Y. 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. Journal of Theoretical Biology 231: 107-120

[25] Ohtsuki H, Iwasa Y. 2006. The leading eight: Social norms that can maintain cooperation by indirect reciprocity. Journal of Theoretical Biology 239: 435-444.

[26] Ohtsuki H, Iwasa Y. 2007. Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. Journal of Theoretical Biology 244: 518-531.

[27] Ohtsuki H, Iwasa Y, Nowak M A. 2009. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature 457: 79-82

[28] Pacheco, J M, Santos, F C and Chalub, F A C C. 2006. Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. PLoS Comp. Biol. 2:1634–1638.

[29] Panchanathan K, Boyd R. 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. Nature 432: 499-502.

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

[31] Rand D G, Ohtsuki H, Nowak MA. 2009. Direct reciprocity with costly punishment: Generous tit-for-tat prevails. Journal of Theoretical Biology 256: 45–57.

[32] Roberts, G 2008. Evolution of direct and indirect reciprocity. Proc. R. Soc. B 275: 173-179

[33] Rockenbach B, Milinski M. 2006. The efficient interaction of indirect reciprocity and costly punishment. Nature 444: 718-723

[34] Rosas A. 2008a. Multilevel selection and human altruism. Biol Philo 23: 205-215.

[35] Rosas A. 2008b. The return of reciprocity. A psychological approach to the evolution of cooperation. Biol Philos 23:555–566.

[36] Scheuring I. 2009. Evolution of generous cooperative norms by cultural group selection. Journal of Theoretical Biology 257: 397–407

[37] Sripada C S, 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

[38] Sugden R. 1986. The Economics of Rights, Cooperation and Welfare. Oxford: Blackwell.

[39] Trivers R. 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46(1): 35-57.

[40] Yamagishi, T. (1986), The Provision of a Sanctioning System as a Public Good, in: Journal of Personality and Social Psychology 51(1): 110-116.