Review TRENDS in Cognitive Sciences Vol.8 No.4 April 2004

Social norms and human cooperation

Ernst Fehr and Urs Fischbacher

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

The existence of social norms is one of the big unsolved problems in social cognitive science. Although no other concept is invoked more frequently in the social sciences, we still know little about how social norms are formed, the forces determining their content, and the cognitive and emotional requirements that enable a species to establish and enforce social norms. In recent years, there has been substantial progress, however, on how cooperation norms are enforced. Here we review evidence showing that sanctions are decisive for norm enforcement, and that they are largely driven by non-selfish motives. Moreover, the explicit study of sanctioning behavior provides instruments for measuring social norms and has also led to deeper insights into the proximate and ultimate forces behind human cooperation.

Human societies represent a spectacular outlier with respect to all other animal species because they are based on large-scale cooperation among genetically unrelated individuals [1]. In most animal societies, cooperation is either orders of magnitude less developed compared with humans, or it is based on substantial genetic relatedness. Cooperation in human societies is mainly based on social norms, including in modern societies, where a considerable amount of cooperation is due to the legal enforcement of rules. Legal enforcement mechanisms cannot function unless they are based on a broad consensus about the normative legitimacy of the rules – in other words, unless the rules are backed by social norms. Moreover, the very existence of legal enforcement institutions is itself a product of prior norms about what constitutes appropriate behaviour. Thus, it is necessary to explain social norms to explain human cooperation.

Social norms are standards of behaviour that are based on widely shared beliefs how individual group members ought to behave in a given situation [2–5]. The group in which social norms prevail can be a family, a peer group, an organization or even a whole society. The group members might obey the norm voluntarily if their individual goals are in line with the normatively required behaviour, or they might be forced to obey the norm because their individual goals differ from the normatively required behavior, in which case the enforcement of the norm presupposes that norm violations are punished. According to a widely shared [6] but not uncontested [7] view, the demand for a social norm arises when actions cause positive or negative side-effects for other people. Environmental pollution, or an individual employee's effort when the team is paid according to the team's total output, are examples that lead to such side-effects. In the team case, a member who contributes to the team's output also raises the pay of the other members. It is therefore in the interest of each individual team member that the other members work hard. This interest in the others' actions is alleged to create the demand for a social norm.

Positive or negative side-effects of individual actions typically give rise to a cooperation or 'public-good' problem. The defining characteristic of a public good, such as clean air or team output, is that no group member can be excluded from the consumption of the good. Therefore, all parties are better off if the public good is provided and the group members share the cost, but each individual also has an economic incentive to free ride; that is, to contribute nothing towards providing the good (see Box 1). This means that if the group members behave according to their economic incentives, they do not cooperate and, therefore, the public good will not be provided. It is clear that a social norm such as '*You should not take advantage of your team members by shirking*' contributes to the provision of the public good.

Box 1. Public goods and Prisoners' Dilemma experiments

By definition, all group members can consume a public good, even those who do not bear the cost of providing the good. Therefore, each member has an incentive to free ride on the contributions of others. In a public good experiment, subjects form groups of $n \geq 2$ individuals and each individual is given a monetary endowment E. Subjects simultaneously decide how much of E they keep for themselves and how much they spend on a group project. The experimenter then multiplies the total amount spent on the group project by a number, denoted by b, that is greater than 1 but smaller than n. The multiplied sum of the member's contribution constitutes the proceeds from the group project. These proceeds are then distributed equally among the n members. This means that each group member, including the contributing subject, earns b/n MUs for every money unit (MU) spent on the project; this is less than 1 because of $b < n$. Yet, the contributing subject has a cost of 1, meaning that a selfish subject never contributes anything to the project in a one-shot experiment. This prediction holds, although it would be collectively rational to contribute everything because if all subjects keep their endowments, they earn E MUs each, whereas if all contribute their endowments the sum of contributions is nE, yielding an income of $(b/n)nE = bE$, which is greater than E for each group member because $b > 1$. For example, if $E = 20$, $b = 2$, subjects earn 20 if nobody contributes and 40 if everybody contributes the whole endowment to the group project.

Formally, the Prisoners' Dilemma (PD) is a special case of the public-good game with $n = 2$ and two available actions: contributing nothing (i.e. defect) or contributing everything (i.e. cooperate). Therefore, each player in the PD is better off if he defects (because $b/n = b/2 < 1$) regardless of what the opponent does.

Corresponding author: Ernst Fehr (efehr@iew.unizh.ch).

 1364-6613/$ - see front matter doi:10.1016/j.tics.2004.02.007

The norm of conditional cooperation

In this article, we review evidence to suggest that human cooperation is largely based on a social norm of *conditional* cooperation. This norm prescribes cooperation if the other group members also cooperate, whereas the defection of others is a legitimate excuse for individual defection. The norm is violated if an individual defects even though the other group members cooperated. The social norm of conditional cooperation provides a proximate mechanism behind the famous tit-for-tat strategy that has been assumed in many evolutionary models [8–10]. To provide evidence for the existence of conditional cooperation, it is first shown that a large percentage of experimental subjects indeed obey the norm. Second, it is shown that norm violators incur punishment.

Evidence for conditional cooperation comes from several experiments in which individuals interact anonymously with each other in a one-shot experiment and where real money is at stake (see Box 1 [11–13]). Typically, a majority of the subjects behaves in a conditionally cooperative manner; that is, they increase their contribution to the public good if the average contribution of the other group members increases (Figure 1). However, there is also a substantial minority of subjects who never contribute anything to the public good. This suggests that discipline of the selfish group members is necessary to enforce widespread cooperation because these members are unwilling to pay for the public good in the absence of sanctions. In addition, it is important to recognize that most subjects who exhibit conditional cooperation are not perfect conditional cooperators. They do cooperate if others also cooperate, but they cooperate less than the others, suggesting that self-interest mitigates adherence to the norm of conditional cooperation. In the absence of sanctions for non-cooperation, this pattern of conditional cooperation – in combination with the existence of complete free riders – is likely to cause decreasing contributions over time. High and stable cooperation levels cannot be maintained if the majority of subjects wants to contribute below the average contribution of the other group members.

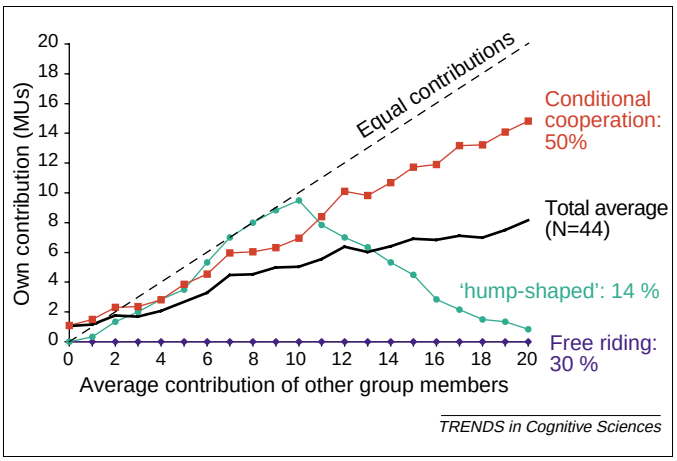

Figure 1. Fischbacher *et al.* [11] elicited subjects' willingness to contribute to a public good, conditional on the average contribution of the other group members. A group consisted of four members and each member could contribute up to 20 'experimental' money units (MUs) to a public good. At the end of the experiment, the MUs were converted into real money according to a publicly known exchange rate (1 MU = $0.30). A selfish member will never contribute anything in this experiment. Despite the economic incentives to contribute nothing, 50% of the subjects raised their contributions if the average contribution of the other members increased (red data plot); 14% exhibited a hump-shaped contribution pattern as a function of others' average contribution (green plot); 30% of the subjects never contributed anything (blue plot); the remaining 6% of the subjects exhibited very irregular contribution patterns. Redrawn from [11] with permission.

The third-party punishment experiment [14], in which a third subject observes the behavior of two subjects who play the Prisoners' Dilemma (PD), is a useful tool for examining whether violations of the conditional cooperation norm are sanctioned. A PD can be viewed as a public goods game with two players in which each player has only two actions: cooperation or defection (see Box 1). If both players defect, the public good is not provided. If both players cooperate, the good is provided and the costs are shared equally, rendering both players better off relative to mutual defection. However, every player has an economic incentive to deviate from mutual cooperation by consuming the public good but letting the other player pay for it.

In the third-party punishment experiment, the two PD players first decided simultaneously whether to cooperate or to defect. Then the third subject, who had a monetary endowment, was informed about both players' actions. Subsequently, the third subject could sanction one, two, or none of the PD players. In reality, sanctions are not costless for the punishing subject. Costs associated with punishing others include the risk of retaliation or at least the potential loss of relationship, the loss of time or money, emotional tensions, and so forth. For this reason, the third party in the experiment incurred a cost if he or she punished a PD player. Each money unit (MU) that the third subject spent on the punishment of a PD player reduced the payoff of the punished PD player by 3 MUs.

Third-party punishment experiments are perfectly suited to study the existence of social norms because the other players' actions do not affect the third subject's economic payoff in any way. The third party is just a passive observer of events that occur in the interaction between other parties. Therefore, the third party has no reason for punishing any of the other players unless a social norm is violated. In fact, because punishment is costly for the third party, and because there are no future benefits from punishing because the experiment is one-shot and the subjects interact anonymously with each other, the desire to punish norm violations has to be strong enough to overcome the third party's self-interest. This means that if punishment by 'disinterested' third parties is observed, one can conclude that there is a strong social norm behind the desire to punish. Third-party punishment experiments have not only been used to study the strength and existence of cooperation norms but also for the study of distribution norms [14–16].

In the PD experiments with third-party punishment, roughly 50% of the subjects in the role of a third party were willing to punish defection of PD players, whereas the punishment of cooperative choices was virtually absent (Box 2). Moreover, defection was punished much more severely if the other PD player cooperated than if the other PD player defected. This indicates that mutual defection is considered much less (if at all) as a norm violation whereas unilateral defection is considered to merit substantial punishment.

Box 2. Third-party punishment

In Fehr and Fischbacher [14], a third party observes the behavior of players in the Prisoners' Dilemma (PD). Then the third party has the option of punishing the PD players. Table I shows the average amount of experimental money (in experimental money units; MU) that an 'impartial' third party spends on the punishment of subjects in the PD. At the end of the experiment, MUs are converted into real money at a known exchange rate (1 MU = $0.37). Subjects who cooperate in the PD are almost never punished by third parties whereas defectors are punished. 45.8% of the third parties ($n = 24$) punish a defector whose partner cooperated. 20.8% punish a defector whose partner also defected. However, the punishment of a defector is much stronger if the other player in the PD cooperated than if the other player in the PD defected indicating a norm of conditional cooperation.

Table I. Number of money units (MUs) spent by a third party in punishing a player in the Prisoners' Dilemma

Punished player is a:	Other player in the PD is a defector	Other player in the PD is a cooperator
Defector	0.583 (20.8%)	3.354 (45.8%)
Cooperator	0.063 (8.3%)	0.083 (4.2%)

Amounts shown are experimental money units. In parentheses is the percentage of third parties who punish.

The impact of sanctioning opportunities on cooperation

What are the conditions under which the norm of conditional cooperation enables groups to establish high and stable cooperation? The existence of a large minority of selfish individuals who violate cooperation norms suggests the need for credible sanctioning threats. In fact, a large amount of evidence [17,18] indicates that stable cooperation is rarely attained in finitely repeated public-goods experiments with anonymous interactions and stable group membership where the selfish choice is full defection. Typically, there is substantial cooperation in the first few periods but over time levels of cooperation decrease, and little cooperation is observed in the final periods (Figure 2).

There are, however, several studies indicating that the addition of sanctioning opportunities and the associated actual sanctioning behaviour has a powerful impact on cooperation rates in these experiments [19–24]. Although the available studies differ in detail, the typical set up is as follows: groups of n individuals (where $n > 2$) are involved in an anonymous public-goods experiment for, say, 10 roughly equal periods of time. In each period, the group members first simultaneously decide how much to contribute. Then all group members are informed about everybody else's contribution (without revealing personal identities), upon which everybody can punish everybody else in the group. Every MU invested into punishment decreases the punished member's monetary payoff by 2–4 MUs. This punishment opportunity usually has a decisive impact on cooperation behaviour. Cooperation decreases if the subjects first conduct an experiment without the opportunity to punish. However, if the same subjects subsequently have the opportunity of punishing, cooperation flourishes (Figure 2). A large increase in cooperation levels already occurs in the very first period after the punishment option is introduced. Moreover, the

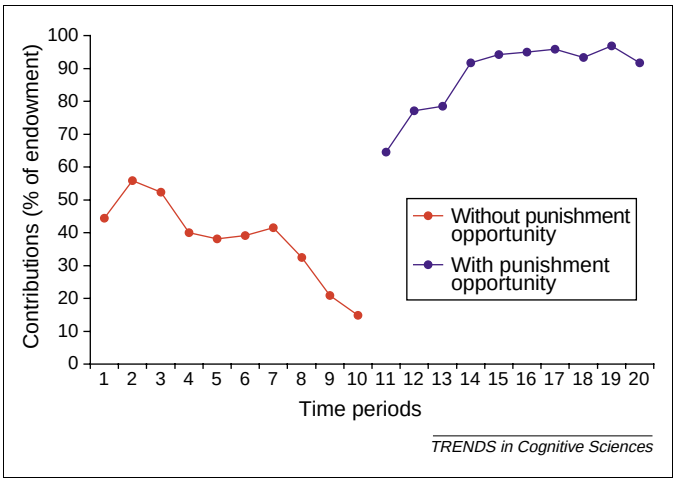

Figure 2. Fehr and Gächter [23] studied the impact of punishment opportunities on cooperation rates in a public-goods experiment. The figure shows subjects' average contributions to the public good (as a percentage of their endowment) over time. During the first ten, roughly equal time periods, no punishment was possible. During periods 11–20, group members could punish each other after they observed each member's contribution level. At the beginning of the experiment cooperation rates of roughly 50% of the endowment were observed, but levels of cooperation decreased over time. The majority of subjects contributed nothing to the public good in period 10, and the rest contributed little. In period 11, the subjects were informed that a new experiment would start in which they would have the opportunity to punish the other group members at a cost to themselves. This modification immediately increased cooperation levels to 65% of the endowment. Then, over time, cooperation rose dramatically,until almost 100% cooperation was attained. Redrawn from [23] with permission.

punishment opportunity not only avoids the decrease in cooperation in the subsequent periods, but even leads to an increase in cooperation. Almost full cooperation is achieved in the final periods. These results are also robust relative to a change in the order of treatments [23]. If the punishment opportunity comes first, cooperation flourishes; cooperation breaks down following its removal.

There are two likely reasons that contribute to the positive impact of the punishment opportunity on cooperation levels. First, a majority of the subjects is willing to punish low contributors; that is, there is a credible punishment threat which disciplines the selfish subjects and the imperfect conditional cooperators (Figure 3). The data also indicate a punishment pattern that is consistent with the norm for conditional cooperation. Those subjects who contribute less than the average contribution of the other group members are heavily punished whereas those who contribute more receive almost no punishment. In addition, the punishment of below-average contributors increases with their deviation from the group average. Second, because the selfish subjects can be disciplined, the norm-abiding subjects can be sure that the other group members will cooperate at high levels. In order words, the punishment opportunity and the associated punishment acts generate a belief that the other group members will cooperate at high levels, and this belief induces the conditional cooperators to cooperate voluntarily at high levels.

Motives behind informal sanctions

No concept in modern game theory presumes that people are selfish. If individual preferences are modeled in game theory, the only requirement is that individuals behave according to consistent goals – irrespective of whether the

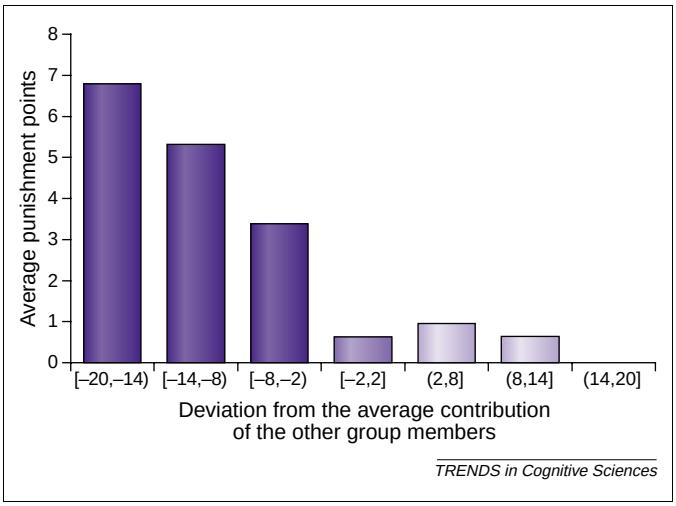

Figure 3. Punishment of group members is conditional on their deviation from the average contribution of the other group members. The horizontal axis indicates the deviation in discrete intervals of an individuals' contribution from the average contribution of the other three group members. The punishment points that are assigned to an individual by the other three group members are shown on the vertical axis. Every member's monetary payoff was reduced by 10% for every punishment point assigned to him. Negative deviations from the other group members' average contribution are strongly punished: a negative deviation in the interval [−20,−14] results in the deviator's income being reduced by almost 70% by the other group members. If an individual's deviation is in the interval [−2,+2] punishment is negligible. Similarly, if an individual deviates positively from others' group average punishment is small. Modified from [23], with permission.

goals are self-regarding or not. Game theory assumes that each actor maximizes utility, given the other actors' choices, but it does not define what matters for utility. In practice, however, the vast majority of game-theory applications assume that *all* people care *only* for their economic self-interest. This literature has led to important theoretical insights into the patterns of interactions and sanctions that are required for the enforcement of cooperation in infinitely repeated public goods situations [25]. According to the self-interest approach, the self-interest of the parties involved drives informal sanctions. However, this approach has never been able to grasp the normative and emotional components of social norms and the resulting actual driving forces behind informal sanctions. The evidence in the third-party punishment experiment contrasts sharply with the predictions of the self interest approach: self-interest cannot drive third-party sanctions because all players involved remain anonymous and there is no future interaction between the players whatsoever.

In principle, the sanctions in the finitely repeated public goods game with a stable group composition (Figure 3) could be driven by self-interest because punished group members typically increase their contributions in future periods. Thus, by punishing low cooperation in period t, the punishing member can benefit from the punished member's higher cooperation in future periods. This possibility for self-interested sanctions is removed if no individual ever interacts with any other individual more than once in the public-goods experiment. In this condition, all sanctions can be attributed to non-selfish motives. This means that if stronger sanctions prevail in a treatment with a stable group composition versus one with a randomly changing group composition (ensuring that no subject ever meets any other subject more than once), there is evidence for selfishly motivated sanctions. The evidence shows, however, that a stable group composition does not generate significantly higher sanctions compared with a treatment with randomly changing group composition (Figure 4) [23,24]. The strength of the sanctions is only slightly and insignificantly higher in the condition with a stable group composition. This result is consistent with questionnaire evidence showing that people's motives for sentencing criminals are found in the area of 'just desserts' rather than deterrence [26]. Thus, in contrast to the conventional assumptions made in game theoretic analyses of sanctioning behaviour, there is little evidence of self-interested sanctions whereas there is much evidence for non-selfishly motivated sanctions.

This raises questions about the motives behind the non-selfish sanctions. Subjects might punish because they want to establish more egalitarian outcomes [27] or because they view unilateral defection as an unfair act that violates the conditional cooperation norm and thus deserves retaliation [28–31]. The evidence suggests that the desire to re-establish equality does not drive the majority of non-selfish sanctions (A. Falk *et al.*, http://www.iew.unizh.ch/wp/iewwp059.pdf). This can be concluded from the fact that subjects punish even in those situations where the sanctions do not change the payoff differences between the punished and the punishing group member. In the case of the sanctions displayed in Figure 4, for example, the costs of the sanctions were the same for the punishing and the punished subject: each MU invested into punishment reduced the payoff of the punished group

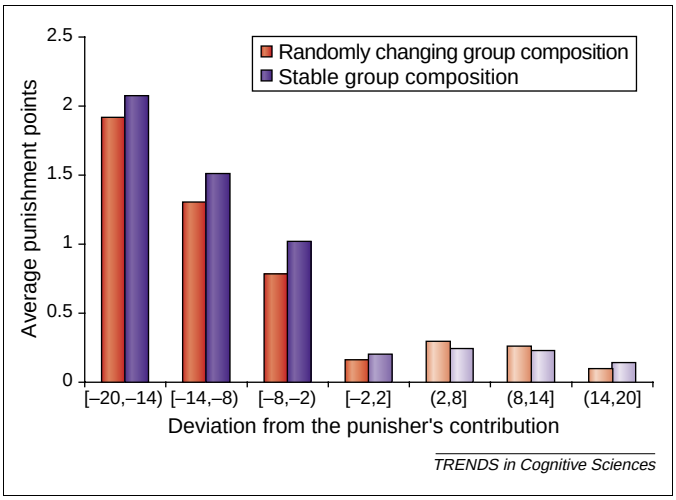

Figure 4. Falk *et al.* studied the impact of self-regarding motives on punishment in a repeated public-goods experiment. The horizontal axis indicates the deviation in discrete intervals of the punished individual's contribution from the punisher's contribution. The vertical axis measures the strength of punishment imposed by individual group members. For every punishment point that an individual assigned to another group member, the punisher and the punished subject had to pay 1 MU. The data shown are based on periods 1–5 of an experiment that lasted six periods. The last period is excluded because punishment can never be motivated by self-regarding motives in the last period. If the group composition is stable, however, a group member can benefit from punishing defectors in the first five periods because the defectors increase their contributions to the public good in the future. If the group composition changes randomly in every period so that no subject meets any other subject more than once, a subject can never reap a selfish benefit from punishing defectors. However, the strength of punishment is roughly the same in both conditions, indicating that selfish motives play little role in the punishment of defectors. Data from Falk *et al.*, in press, http://www.iew.unizh.ch/wp/iewwp059.pdf.

member by exactly 1 MU. Despite this 'unfavorable' punishment technology, punishment was still wide spread and strong.

The absence of evidence for selfishly motivated sanctions does not mean that self-interest plays no role at all in sanctioning behavior. From an economic viewpoint, one would expect that as the costs of sanctions increase for the sanctioning subject, there will be fewer and weaker sanctions. Recent research has shown that this is indeed the case (C.M. Anderson and L. Putterman, unpublished; J.P. Carpenter, unpublished). Thus, as the sanctioning subjects have to give up more of their own payoff to sanction others, they punish less, indicating that they trade off their own economic payoff with the non-pecuniary 'benefits' of sanctioning others. This fact is also important with respect to how one views and analyzes social norms. It is apparently wrong to assume that social norms are followed or enforced regardless of the cost of norm adherence and norm enforcement. If norm adherence or norm enforcement becomes more costly, norms are more likely to break down. From a methodological viewpoint this also means that norms can be analyzed by suitably adjusted game theoretic models. These models assume that people are not only motivated by economic self-interest but also by norms of fairness and reciprocity [27–33]. Recent studies on the neurobiology of cooperation [34] and punishment [35] are consistent with these models.

The evolution of human cooperation

The human capacity to establish and enforce social norms is perhaps the decisive reason for the uniqueness of human cooperation in the animal world [36]. The evidence indicates that other animals largely lack the cognitive and emotional capacities that are necessary for social norms [37,38]. In comparison with humans, most animals have a very high rate of time discounting, lack the ability of precise numerical discrimination, exhibit serious memory constraints or lack inhibitory control [38]. Therefore, powerful ultimate forces, that probably had little impact on animal evolution [38,39], supported the evolution of cooperation in humans. Reciprocal altruism [8], reputation-based altruism [40,41], and punishment-based altruism [42] are much more likely to generate successful cooperation norms if the actors are less impatient, exhibit inhibitory controls and are less constrained by memory limits. In addition, cultural group selection is also likely to be more important among humans because of human cognitive capacities [42–46]. The rationale for this is simple and can be illustrated in the context of the public-goods experiments. If there is a punishment opportunity, as was likely the case in hunter gatherer groups, selfish types do not flourish because they are sanctioned. Social norms and the associated sanctions remove the within group selection advantage of norm violators and favor the selection of norm-abiding behaviors within the groups. There is abundant anthropological evidence indicating that human groups differ greatly in their social norms [47,48]. Thus, the existence of social norms creates conformity within groups and heterogeneity across groups. This provides the raw material for cultural group selection to become effective because it requires the existence of persistent differences between groups plus the plausible assumption that groups with more cooperative norms are more likely to prevail in group conflicts.

Conclusions and outstanding questions

It is not possible to understand the peculiarities and the forces behind human cooperation unless we understand social norms. Experimental evidence indicates the existence of a norm of conditional cooperation. If other group members cooperate, the norm also requires us to cooperate; if others defect we are also allowed to defect. 'Disinterested' third parties frequently punish violations of the conditional cooperation norm for non-selfish reasons. More generally, non-selfishly motivated punishment constitutes a powerful device for the enforcement of social norms and human cooperation. Despite some recent progress in the analysis of social norms [48–51], a long journey still remains until science will be able to provide a satisfactory understanding. Although it is known that emotions are likely to play a key role in cooperation [34] and punishment decisions [24,35], we still do not know whether they drive these decisions or whether they are merely associated with these decisions. A related problem is that the neural underpinnings of social norms are still largely uninvestigated. There is also a decisive lack of knowledge regarding the social and the economic determinants of social norms. The socio-economic environment shapes the costs and benefits of cooperation and punishment and is thus likely to be an important determinant of social norms but empirically as well as theoretically we still know little about this. More knowledge in this area is likely to help us to understand which environment is likely to favor which norms and thus, when norms are stable and when they will change.

Acknowledgements

We gratefully acknowledge support by the Ludwig Boltzmann Institute for the Analysis of Economic Growth, by the Swiss National Science Foundation (project no. 12–67751.02) and by the MacArthur Foundation Network on Economic Environments and the Evolution of Individual Preferences and Social Norms.

References

1 Henrich, J. *et al.* (2003) The cultural and genetic evolution of human cooperation. In *Genetic and Cultural Evolution of Cooperation* (Hammerstein, P., ed.), pp. 445–468, MIT Press
2 Elster, J. (1989) *The Cement of Society: A Study of Social Order*, Cambridge University Press
3 Horne, C. (2001) Sociological perspectives on the emergence of norms. In *Social Norms* (Hechter, M. and Opp, K.D., eds), pp. 3–34, Russell Sage Foundation
4 Ellickson, R.C. (2001) The evolution of social norms: a perspective from the legal academy. In *Social Norms* (Hechter, M. and Opp, K.D., eds), pp. 35–75, Russell Sage Foundation
5 Voss, T. (2001) Game theoretical perspectives on the emergence of social norms. In *Social Norms* (Hechter, M. and Opp, K.D., eds), pp. 105–138, Russell Sage Foundation
6 Coleman, J. (1990) *Foundations of Social Theory*, Belknap Press/ Harvard University Press
7 Elster, J. (1989) Social norms and economic theory. *J. Econ. Perspectives* 3, 99–117
8 Axelrod, R. and Hamilton, W.D. (1981) The evolution of cooperation. *Science* 211, 1390–1396
9 Boyd, R. and Richerson, P.J. (1988) The evolution of reciprocity in sizable groups. *J. Theor. Biol.* 132, 337–356

10 Nowak, M. and Sigmund, K. (1993) A strategy of win–stay, lose–shift that outperforms tit-for-tat in the Prisoners' Dilemma game. *Nature* 364, 56–58
11 Fischbacher, U. *et al.* (2001) Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* 71, 397–404
12 Keser, C. and van Winden, F. (2000) Conditional cooperation and voluntary contributions to public goods. *Scand. J. Econ.* 102, 23–39
13 Brandts, J. and Schram, A. (2001) Cooperation and noise in public goods experiments: applying the contribution function approach. *J. Pub. Econ.* 79, 399–427
14 Fehr, E. and Fischbacher, U. Third party sanctions and social norms. *Evol. Hum. Behav.* (in press)
15 Kahneman, D. *et al.* (1986) Fairness as a constraint on profit seeking: entitlements in the market. *Am. Econ. Rev.* 76, 728–741
16 Turillo, C.J. *et al.* (2002) Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. *Org. Behav. Hum. Dec. Proc.* 89, 839–865
17 Dawes, R. and Thaler, R. (1988) Cooperation. *J. Econ. Perspect.* 2, 187–197
18 Ledyard, J. (1995) Public goods: a survey of experimental research. In *Handbook of Experimental Economics* (Kagel, J. and Roth, A., eds), pp. 111–194, Princeton University Press
19 Ostrom, E. *et al.* (1992) Covenants with and without a sword – self-governance is possible. *Am. Polit. Sci. Rev.* 86, 404–417
20 Ostrom, E. (2000) Collective action and the evolution of social norms. *J. Econ. Perspect.* 14, 137–158
21 Yamagishi, T. (1986) The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* 51, 110–116
22 Yamagishi, T. (1988) Seriousness of social dilemmas and the provision of a sanctioning system. *Soc. Psych. Q.* 51, 32–42
23 Fehr, E. and Gächter, S. (2000) Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* 90, 980–994
24 Fehr, E. and Gächter, S. (2002) Altruistic punishment in humans. *Nature* 415, 137–140
25 Fudenberg, D. and Maskin, E. (1986) The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54, 533–556
26 Carlsmith, K.M. and Darley, J.M. (2002) Why do we punish? Deterrence and just desert motives for punishment. *J. Pers. Soc. Psych.* 83, 284–299
27 Fehr, E. and Schmidt, K.M. (1999) A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868
28 Rabin, M. (1993) Incorporating fairness into game theory and economics. *Am. Econ. Rev.* 83, 1281–1302
29 Falk, A. and Fischbacher, U. (2001) Distributional consequences and intentions in a model of reciprocity. *Ann. Econ. Stat.* 63, 111–129
30 Dufwenberg, M. and Kirchsteiger, G. A theory of sequential reciprocity. *Games Econ. Behav.* (in press)
31 Levine, D.K. (1998) Modeling altruism and spitefulness in experiments. *Rev. Econ. Dynam.* 1, 593–622
32 Lindbeck, A. *et al.* (1999) Social norms and economic incentives in the welfare state. *Q. J. Econ.* 114, 1–35
33 Dufwenberg, M. and Lundholm, M. (2001) Social norms and moral hazard. *Econ. J.* 111, 506–525
34 Rilling, J.K. *et al.* (2002) A neural basis for social cooperation. *Neuron* 35, 395–405
35 Sanfey, A.G. *et al.* (2003) The neural basis of economic decision-making in the ultimatum game. *Science* 300, 1755–1758
36 Fehr, E. and Fischbacher, U. (2003) The nature of human altruism. *Nature* 425, 785–791
37 Stephens, D.W. *et al.* (2002) Discounting and reciprocity in an iterated Prisoner's Dilemma. *Science* 298, 2216–2218
38 Stevens, J.R. and Hauser, M.D. (2004) Why be nice? Psychological constraints on the evolution of cooperation. *Trends Cogn. Sci.* 8, 60–65
39 Hammerstein, P. (2003) Why is reciprocity so rare in social animals? In *Genetic and Cultural Evolution of Cooperation. Dahlem Workshop Report 90* (Hammerstein, P., ed.), MIT Press
40 Nowak, M.A. and Sigmund, K. (1998) Evolution of indirect reciprocity by image scoring. *Nature* 393, 573–577
41 Gintis, H. *et al.* (2001) Costly signaling and cooperation. *J. Theor. Biol.* 213, 103–119
42 Boyd, R. and Richerson, P.J. (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* 13, 171–195
43 Bowles, S. *et al.* (2003) The coevolution of individual behaviors and group level institutions. *J. Theor. Biol.* 223, 135–147
44 Boyd, R. *et al.* (2003) The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3531–3535
45 Gintis, H. (2000) Strong reciprocity and human sociality. *J. Theor. Biol.* 206, 169–179
46 Henrich, J. and Boyd, R. (2001) Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *J. Theor. Biol.* 208, 79–89
47 Sober, E. and Wilson, D.S. (1998) *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Harvard University Press
48 Henrich, J. *et al.* (2001) In search of homo economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* 91, 73–78
49 Hechter, M. and Opp, K.D. (2001) *Social Norms*, Russell Sage Foundation
50 Posner, E.A. (2000) *Law and Social Norms*, Harvard University Press
51 Hechter, M. and Opp, K.D. (2001) What have we learned about the emergence of social norms. In *Social Norms* (Hechter, M. and Opp, K.D., eds), Russell Sage Foundation