

Social Norms and Reciprocity*

Andreas Diekmann
Institut für Soziologie
Universität Bern

Thomas Voss
Institut für Soziologie
Universität Leipzig

[March 2003]

Paper presented on the session „Solidarity and social norms – models and mechanisms“,
Sektion Modellbildung und Simulation, Leipzig, October 2002

* Manuela Vieth detected an error in an earlier version of this paper. We would like to thank her very much for the highly valuable comments. Thomas Voss' work on this paper has been supported by a fellowship of the Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS) in the academic year 2001/2002. Research grant Vo 684/5-1 of Deutsche Forschungsgemeinschaft is also gratefully acknowledged.

1. Introduction

Informal social norms with sanctions can be explained, in principle, by a rational choice approach. Coleman (1990) and others (Ellickson 1991; see also Hechter & Opp 2001) have argued that norms serve to improve the efficiency or the aggregate welfare of the norm beneficiaries. For example, the farmers studied in Ellickson's seminal work (1991) have privately (i.e. without the support of third parties besides the community of farmers) enforced norms that foster cooperation among neighbors. Such norms govern the solution of conflicts in cases of cattle trespassing or provide rules on how to share costs of constructing boundary fences in a large agricultural area. It is commonly argued that a *demand* for norms arises in situations of a social dilemma, such as the prisoner's dilemma (e.g., Ellickson 1991: Chapter 9). Norms are, however, effectively realized only if there are sanctions that punish non-cooperative behavior. Where do these sanctions come from? An obvious answer is that *repeated interactions* among norm beneficiaries help to enforce norms of cooperation. Given repeated interactions, strategies of conditional cooperation will be possible. These strategies contain a sanctioning mechanism that triggers defection if a partner defects. The logic of repeated games is the rationale of Ellickson's hypothesis on the effects of a "close-knit community" on the enforcement of norms. In a similar vein, Coleman (1990: Chapter 11) has argued that the existence of a closed network of social relations is important for the realization of norms. This paper does not deny the importance of repeated interactions in the context of norms. However, two points that are neglected in the traditional approach of the Coleman-Ellickson type, will be addressed: First, it is somewhat misleading to study norms only by means of using stage games of the prisoner's dilemma kind. This is so because repetitions of the prisoner's dilemma give rise to *indirect* sanctions of using the Nash threat to defect. Ethnographic evidences abound that in real life many other, more active, types of sanctions are used: ostracism, physical retaliation, refusal of social approval, gossip etc. It is therefore important to study games that represent situations where the actors have an opportunity to apply *direct* sanctions. A very simple case of such a game is the *norm game* (Voss 2001). This is a *modified* prisoner's dilemma with an additional punishment phase. Secondly, given a norm game with a punishment option, the question of the possibility of cooperation in the *one-shot* game seems appropriate because the norm game is strategically different from the prisoner's dilemma. In a norm game, under certain conditions, there exist Nash equilibria of mutual cooperation. Experimental work (Fehr & Gächter 2000b) demonstrates that even in one-shot situations the level and proportion of cooperative behavior increases if an punishment option is available to the players of a public goods game. It is therefore important to analyze conditions such that this is consistent with a rational choice approach. The paper is meant as a first step toward this task. The main result will be that non-standard assumptions about human motivations or preferences can explain norms with sanctions even in one-shot situations. This is shown by an analysis of the norm game with two well-known recent models of fairness (Bolton & Ockenfels 2000; Fehr & Schmidt 1999) from behavioral game theory.

2. Cooperation in a norm game

The norm game is a two-person non-cooperative game with the following properties (Voss 2001): There are two phases of the game. The first comprises an ordinary prisoner's dilemma with two options, namely cooperation and defection. In the second phase, both players can react on the decisions of the first stage by punishing their partner. The payoffs are simply assumed to be sums of the payoffs that are collected in both phases. With respect to the prisoner's dilemma it is, for convenience, assumed that for both players $T > R > P > S$. In the punishment phase, a player is able to sanction her partner. The target's cost or fine of being (passively) punished is denoted by p . The cost of (actively) punishing is k . Is it possible, under assumptions of individual rationality, that both actors comply with the norm? Rationality in this context means that compliance with the norm (cooperation) is (i) a Nash equilibrium strategy and (ii) the equilibrium is subgame perfect (spe). Subgame perfection is an important rationality criterion because the norm game is sequential, that is, the players are able to use *threats* to punish defections. These threats need to be credible, if standards of game theoretical rationality are used. In other words, in sequences of the game which are "out of equilibrium" there should be incentives to apply those punishments.

It is easily seen (Voss 2001) that there exists a Nash equilibrium that yields mutual cooperation if $p \geq T - R$. In other words, cooperation is possible if the fine from the partner's punishment is at least as large as the costs of cooperation ($T - R$). Intuitively, this means that the prisoner's dilemma is transformed into a different game, namely a kind of assurance game, if sanctions are possible. Another obvious result, however, is that this equilibrium is subgame perfect under the condition that $k \leq 0$. In other words, mutual cooperation which is based on reciprocal threats to punish defections in the second phase is possible only under the condition of costless sanctioning. If punishment is costly to the agent who actively sanctions his partner, that is, if $k > 0$, the threat to use it will not be credible in the one-shot case. (N.B.: This is different in the repeated game, but we will not dwell on that point here.) The game theoretic analysis of the norm game therefore leads to the following conclusion with respect to the so-called second order dilemma (sanctions yield norm conformity, but there is free riding with regard to the costs of using sanctions): *In a one-shot norm game no subgame perfect equilibrium of mutual cooperation exists if sanctions are costly ($k > 0$).*

3. Non-standard preferences in the one-shot norm game

Besides anecdotal evidence there is experimental work that is relevant to actual behavior in the one-shot norm game. Fehr and Gächter (2000b, 2002) run public goods experiments with an additional option of punishing. Surprisingly, there were substantial effects of the punishment option even in a one-shot case: If punishment is *possible*, the level of cooperation with regard to the public good is higher. Furthermore, costly punishment was *in fact applied* to sanction free riders. These results are important to our discussion of norms because the public goods game of Fehr and Gächter is structurally similar to the norm game. We expect that analogous experimental results would be obtained with regard to the simpler, two person norm game.

Section 2 has reported the analytical result that *no* subgame perfect equilibria of norm conformity are possible in the one-shot case. We now ask whether a game theoretic approach that uses assumptions on *non-standard motivations* (preferences) can explain cooperation and thus norm conformity in this context.

The assumption that individuals' preferences are more than functions of the material payoffs is of course quite common in the tradition of individualist social theory. Alluding to Adam Smith's theory of moral sentiments, Frank (1988) argued that human actors are disposed to show certain *emotional* reactions if their partners cheat them or deviate from moral norms. Reasoning along these lines one is tempted to conclude that the function of certain emotional dispositions is to make threats or promises credible. For example, an emotional disposition that may be called "*negative reciprocity*" (Fehr and Gächter 2000a) yields costly punishments as a reaction toward a partner who defected in some sense even in *one-shot situations*. Given this, under the assumption of complete information about these preferences or dispositions, the partner who anticipates this reaction rationally cooperates in a one-shot situation, provided the reciprocity motive is strong enough.

Behavioral and evolutionary game theory recently started to develop models of rational (or boundedly rational or at least adaptive) behavior such that the standard assumptions of a homo economicus are dispensed with. Whereas homo economicus is self-regarding and outcome-regarding, it seems more realistic to assume a model of man such that preferences are also other-regarding and process-regarding (Gintis 2000: 251; cf. Gintis 2000: Chapter 11 for an excellent survey on these new models).

Two models from behavioral game theory seem particularly suitable to analyze other-regarding behavior. The *ERC* (Equity, Reciprocity, and Competition) model (Bolton and Ockenfels 2000) and the Fehr and Schmidt (*FS*) model (Fehr and Schmidt 1999) represent von Neumann-Morgenstern utility functions with non-standard preferences that include a "social component". The arguments of the utility function are *material outcomes*. We assume $T > R > P > S > 0$. In both models, utility functions are depending on two kinds of arguments. There is first a component f that depends on an individual i 's material payoffs y_i . A second term g reflects the deviation of the material payoff from a reference payoff. In both models, individuals experience disutilities if there inequalities between material outcomes:

(Inequality Aversion) $u(y_i) = f(y_i) - g(\text{deviation of } y_i \text{ from reference payoff})$.

The social component of ERC preferences is defined such that actors compare their own material payoff with the *average* outcome in the reference group. There is a behavioral tendency to be inequality averse: Absolute utility decreases if an actor receives less (or more) than the reference group. The ERC models is a relatively general theoretical tool that can deal with a large set of games. It is given in an axiomatic form. Depending on specifications of the parameters of the utility functions, it reflects different degrees of egoism or fairness (or a population with heterogeneity). Completely selfish behavior is a limiting case of the model. We use a special simplified version of ERC for 2 person games ($i=1,2$):

$$(ERC) u_i = a_i y_i - b_i (0.5 - \sigma_i)^2$$

In (ERC) i 's utility is depending on i 's material outcomes y_i and on a reference standard σ_i that reflects the share of i 's payoffs relative to the sum of the payoffs of the group members. In this specification, it is assumed that $n = 2$. Hence, the fair share is 0.5. If $\sigma_i = 0.5$, the actor's behavior will correspond to usual positively linearly increasing utility in money (or in other cardinal material payoffs). The parameters a_i and b_i (with $a_i > 0$; $b_i > 0$) represent the amounts of selfish or other-regarding orientations, respectively. If $b_i \rightarrow 0$, the actor's preferences are, in the limiting case, completely self-regarding. The exponent of the second term represents the

idea that larger differences from the fair allocation nonlinearly affect (dis)utility. Notice that in the specification of ERC that is used here it is assumed that one feels equally uncomfortable if one receives too much or too little compared with the standard. This is, however, not a necessary assumption, but is employed merely for convenience and can in principle be dispensed with.

In contrast to ERC, the Fehr and Schmidt (FS) model does not compare an individual's share with the (average) group outcome, but the differences of an actor i 's outcome and each other actor's outcomes are summed over all persons who are "richer" or "poorer" than i , respectively. For the special case of a two-person game the utility function is:

$$(FS) \quad u_i = y_i - a_i \max\{y_j - y_i, 0\} - b_i \max\{y_i - y_j, 0\}$$

The parameters a_i and b_i are individual-specific. It is assumed that $0 \leq b_i < 1$; $b_i \leq a_i$.

In the following, we present results of an analysis of the one-shot norm game that is based on these assumptions:

(A1) The players who are involved in a norm game are completely rational but are endowed with non-standard preferences of the ERC- or FS-type.

(A2) The players are acting under complete information with regard to each other's preferences.

Proposition 1 (ERC): Given ERC preferences, under certain conditions there exists a pair of cooperative strategies in the one-shot norm game which are subgame perfect equilibrium strategies.

Proposition 2 (ERC): Given ERC preferences, a necessary although not sufficient condition for a pair of strategies in the one-shot norm game to yield cooperation as a subgame perfect equilibrium is, for $k > 0$, that $p/k \geq T/S$.

The proof of propositions 1 and 2 follows from the fact that subgame perfection requires that in the second phase of the norm game $u(s^*) \leq u(s)$. Here $u(s^*)$ denotes the utility without sanctioning and $u(s)$ is the utility of an actor sanctioning the defective co-player. In other words, after the partner's defection (which is albeit "out of equilibrium") there must be no positive incentive to deviate from the application of a punishment (s).

Proposition 3 (FS): Given FS preferences, a necessary and sufficient condition for a pair of strategies in the one-shot norm game to yield cooperation as a subgame equilibrium is, for $k > 0$, that $ap \geq k(1 + a)$.

Notice that in the case of $T-p > S-k$, such that the defector gains more than the cooperator, subgame perfection requires $u(s^*) \leq u(s)$ with

$$\begin{aligned} u(s^*) &= S - a_i(T-S), \\ u(s) &= S - k - a_i((T-p)-(S-k)). \end{aligned}$$

Hence, $p/k \geq (1+a_i)/a_i$.

4. Conclusions

The paper has shown that rational actors are able to enforce social norms with sanctions even in one-shot situations and for costly punishments ($k > 0$). The result can be obtained by using two recent models of non-standard preferences, namely the ERC- and the FS-models. In both cases, there are conditions such that a behavioral tendency to avoid inequalities enables actors to make threats to punish defections credible. Notice that this is not an ad hoc explanation for cooperation in a norm game. ERC- and FS-models are relatively general theoretical instruments to model non-standard motivation functions in a variety of circumstances. They were not designed to analyze norm games. Therefore, our results, in principle, contribute to an enlargement of the empirical content of these models.

With regard to the substantive problem of explaining norms, it is important to note that our paper does argue that the second order problem can be resolved in the one-shot case if certain non-standard motivations are assumed. We do not argue against individual rationality and rational choice. In our interpretation, other-regarding preferences are commitment devices that transform negative reciprocity (and possibly also positive reciprocity) into a rational strategy that is consistent with basic rationality criteria of standard game theory.

References

- Bolton, Gary and Axel Ockenfels 2000: ERC – A theory of equity, reciprocity, and competition, *American Economic Review* 90: 167-193
- Coleman, James S. 1990: *Foundations of Social Theory*, Cambridge, Mass.
- Ellickson, Robert C. 1991: *Order without Law*, Cambridge, Mass.
- Fehr, Ernst and Klaus Schmidt 1999: A theory of fairness, competition, and cooperation, *Quarterly Journal of Economics* 114: 817-868
- Fehr, Ernst and Simon Gächter 2000a: Fairness and retaliation: A theory of reciprocity, *Journal of Economic Perspectives* 14: 159-181
- Fehr, Ernst and Simon Gächter 2000b: Cooperation and punishment in public goods experiments, *American Economic Review* 90: 980-994
- Fehr, Ernst and Simon Gächter 2002: Altruistic punishment in humans, *Nature* 415: 137-140
- Gintis, Herbert 2000: *Game Theory Evolving*, Princeton, N.J.
- Hechter, Michael and Karl-Dieter Opp (Eds.) 2001: *Social Norms*, New York
- Voss, Thomas 2001: Game theoretical perspectives on the emergence of social norms, pp. 105-136 in: Michael Hechter and Karl-Dieter Opp (eds.), *Social Norms*, New York