ECONOMICS BULLETIN

A trace of anger is enough: on the enforcement of social norms

Jakub Steiner
The University of Edinburgh

Abstract

It is well documented that the possibility of punishing free-riders increases contributions in one-shot public good games. I demonstrate theoretically that minimal punishment commitments (perhaps provided by anger) may lead to high contribution levels. Thus, almost selfish players may behave as strong reciprocators.

I benefited from the comments of Ken Binmore, Dirk Engelmann, Andreas Ortmann, Avner Shaked, and seminar participants at Max Planck Institute in Jena. Laura Strakova carefully edited the manuscript. The work was supported by a CERGE-EI / World Bank grant. Author's second affiliation is CERGE-EI in Prague.
Citation: Steiner, Jakub, (2007) "A trace of anger is enough: on the enforcement of social norms." *Economics Bulletin,* Vol. 8, No. 1 pp. 1-4
Submitted: January 19, 2007. **Accepted:** January 19, 2007.
URL: http://economicsbulletin.vanderbilt.edu/2007/volume8/EB-07H40001A.pdf

1 Introduction

Experimental evidence suggests that people increase their contributions to a public good substantially when the possibility of punishing free-riders is introduced (e.g. Fehr and Gächter, 2000a, 2002; Yamagishi, 1986). Since costly punishment is not subgame perfect, this evidence is often interpreted as another blow to the concept of *homo oeconomicus* and as evidence supporting strong reciprocity theory. Indeed, some deviation from the *homo oeconomicus* assumptions is needed to explain the (strong) consequences of the possibility to punish. This deviation is typically interpreted as emotions, particularly anger, in the reciprocity literature (e.g. Fehr and Gächter, 2002). Below, I theoretically demonstrate that only traces of anger are needed to explain the well-documented high contribution levels; the deviation thus needs to be rather small. This short note is an instance of a more general program: small deviations from the *homo oeconomicus* framework explain many large behavioral anomalies documented in behavioral experiments, see Binmore (2006) and Binmore and Swierzbinski (2006).

The model under study reflects the standard experimental design of public good games with punishment option: N players have a possibility to contribute to a linear public good. All players are thereafter given the opportunity to allocate punishment points to other players, which is costly for both sides.

Though I assume that the punishment activity is motivated by emotions such as anger, it is not the purpose of this note to explain where emotions come from.[1] The focus is on the intensity of emotions that it takes to enable the players to cooperate, where the intensity is measured by an amount of money the players spend on punishment. I choose a black box approach, and model punishment as an automatic reaction fully governed by a punishment rule, which is a function of the realized contributions.

Below I construct a punishment rule which induces a game with *unique* equilibrium in which all players make high contributions. The rule requires every player to punish only the player with the lowest contribution which in turn motivates her to escape the lowest position. The game with such a punishment rule can be solved by iterated elimination of dominated strategies because the lowest non-eliminated contribution level is always dominated by the level just above it. At the same time, it is possible to implement such a punishment rule at minimal cost because resources from a large group of players are focused on one free-rider. The rule resembles the rule experimentally implemented by Yamagishi (1986). I do not argue that this exact punishment rule is necessarily used in reality; I simply show that a rule inducing high contributions exists.

[1] People may punish a bit because of warm glow (e.g. Andreoni, 1990), altruism (e.g. Andreoni and Miller, 2002), inequality aversion (e.g. Bolton and Ockenfels, 2000;), or reciprocity (e.g. Levine, 1998). The Role of emotions as a commitment device was stressed by Frank (1988). There is neuroscientific evidence suggesting that emotions such as anger are triggered by unfair behavior (de Quervain et al., 2004).

2 The Model

Each player $i \in \{1, \ldots, N\}$ chooses a contribution level c^i from the common strategy set $S = \{\frac{0}{L}\bar{c}, \frac{1}{L}\bar{c}, \ldots, \frac{L}{L}\bar{c}\}$, where L represents the number of levels approximating the continuous interval $[0, \bar{c}]$, and L is assumed to be large. The maximal possible contribution is $\bar{c}$. After the contributions $\mathbf{c} = (c^1, \ldots, c^N)$ of all players are realized, and observed by everyone, the players automatically assign punishment points to each other. I abstract from the individual punishment actions and analyze only the sum of punishments from all players, which is denoted as the aggregate punishment $p^i(\mathbf{c}) \geq 0$ of player i. The unspecified punishment rules of individual players may in reality be heterogenous (e.g. Fehr and Gächter, 2000b), and the aggregate punishment rule $p^i(\cdot)$ can be interpreted as the expected punishment of player i given the population the players are drawn from.

Although the game has two time phases, it can be seen conceptually as a one-stage game as the players make decisions only in the contribution phase. The automatic punishment can be modeled within the payoff function of the one-stage game.

Definition 1. *The game of N players with strategy sets $\{S^i = S\}_{i=1}^N$ and the payoff function*[2]

$$U^i(\mathbf{c}) = -c^i - p^i(\mathbf{c}) \tag{1}$$

is the punishment game.

The central assumption is that the average player is able to commit to spend only a penny to punish free-riders:

A1: $\frac{\sum_{i=1}^N p^i}{N} \leq 1$.

The set of equilibria of the punishment game depends on the particular punishment rule. With the assumption **A1** being valid, one might expect an unattractive equilibrium in which all players free-ride and punishment is ineffectively spread among all the players resulting in no player having an incentive to contribute. However, the punishment rule proposed below induces a game with a unique equilibrium in which all participants make high contributions:

Denote the lowest contribution among players by c_l and the second lowest by c_{sl} where $c_l = c_{sl}$ if there is more than one player with the lowest contribution. Let the punishment rule be

$$p^i(\mathbf{c}) = \begin{cases} 0 & \text{if } c^i > c_l \text{ or } c^i = \bar{c}, \\ \frac{N}{\bar{c} + \frac{\bar{c}}{L}}(c_{sl} + \frac{\bar{c}}{L} - c^i) & \text{if } c^i = c_l. \end{cases} \tag{2}$$

[2] The payoffs from the public good are the same for all players; therefore, they do not have to be included in the payoff function; c^i is interpreted as contribution cost net of the increase of the public good. The small cost of punishment is also not included because punishment is modeled as an automatic action not as a result of payoff optimization.

This punishment rule motivates the player(s) with the lowest contribution to increase her(their) contribution(s) one level above the second lowest contribution. The marginal punishment $\frac{N}{\overline{c}+\frac{\overline{c}}{L}}$ is chosen such that the total punishment expenditures are always at most 1 unit per player.

Proposition 1. 1. *The punishment game of N players with punishment rule (2), with the maximal possible contribution satisfying $\overline{c}+\frac{\overline{c}}{L} < N$, and with a large number of contribution levels $L > N$ has a unique equilibrium with all players contributing the maximal possible amount $\{c^{*i} = \overline{c}\}_{i=1}^{N}$.*

2. *The punishment rule (2) satisfies the low cost assumption* **A1**.

Proof. 1. The player with the lowest contribution always wishes to increase her contribution by at least $\frac{\overline{c}}{L}$ because the marginal punishment she experiences $\frac{N}{\overline{c}+\frac{\overline{c}}{L}} > 1$ is higher than the marginal cost of contribution. Hence, the lowest contribution level, 0, is dominated by $\frac{\overline{c}}{L}$. After elimination of $\{0, \frac{1}{L}\overline{c}, \ldots, \frac{k}{L}\overline{c}\}$, $\frac{k+1}{L}\overline{c}$ is dominated by $\frac{k+2}{L}\overline{c}$ because $\frac{k+1}{L}\overline{c}$ would be the lowest contribution among non-eliminated strategies, for $k = 1, \ldots, L-2$. Thus, the game can be solved by iterated elimination of dominated strategies; only $\overline{c}$ survives this process.

2. There is either only one player with the lowest contribution, and then she is the only one being punished. The punishment is largest in this case if $c_{sl} = \overline{c}$ and $c_l = 0$. Then the punishment is $\frac{N}{\overline{c}+\frac{\overline{c}}{L}}(\overline{c}+\frac{\overline{c}}{L} - 0) = N$, so the cost is at most 1 unit per player.

Or there may be many players with the lowest contribution, but then $c_{sl} = c_l$ so each punishment is $\frac{N}{\overline{c}+\frac{\overline{c}}{L}}\frac{\overline{c}}{L} \leq 1$, thus the cost per player is smaller than 1 unit. □

The punishment rule (2) specifies that players punish each other even after they all have colluded on a same contribution level. Steiner (2006) analyzes a related model in which players cannot punish other players who contributed the same amount as themselves. Naturally, each contribution level is a Nash equilibrium under this additional demanding assumption. However, there exists a rule, related to (2) which guarantees that only the highest contribution survives stochastic evolution in a sense of Ellison (2000).

3 Concluding Remarks

The possibility to punish free-riders acts like a magnifying glass on the players' emotionality. Although they give up only a penny voluntarily, they are able to induce each other to contribute N pennies. Thus, the cooperation level increases linearly with the number of players. This insight is confirmed by Carpenter (2004) who experimentally found a positive group size effect even if one controls for marginal group return of contributions.

The punishment rule proposed here has a normative appeal. If a group of subjects wants to overcome the free-riding problem but has only a limited possibility to commit to a punishment threat, it is this rule they should choose. Of course, the punishment game requires quite a bit of information: all players have to be able to monitor actions of all other players. This is realistically possible only in small groups, such as workplaces.

Sometimes the public good will be too expensive to be provided through the punishment game discussed above. However, the game can be extended in a straightforward manner to a second punishment stage where those who did not participate enough on punishing free-riders could be punished. Then the players could enforce N^2 contribution levels.

The model proposed here demonstrates that punishment games may be very efficient institutions even for high pecuniary stakes when emotional payoffs are small in the comparison. It also suggests that experimental results on the punishment game do not have to be interpreted as clear refutation of the selfishness assumption: Only minimal deviation from the assumption is needed to explain cooperation in the punishment game. Strong reciprocity is strong in its consequences, not in the emotional requirements of players.

References

[1] Andreoni J., 1990, Impure Altruism and Donation to Public Goods: A Theory of Warm-Glow Giving?, The Economic Journal 100, 464–477.

[2] Andreoni J. and J.H. Miller, 2002, Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism, Econometrica 70, 737–753.

[3] Binmore K., 2006, Why do people cooperate?, Politics, Philosophy & Economics 5, 81–96.

[4] Binmore K. and J. Swierzbinski, 2006, A Little Behavioralism Can Go a Long Way, mimeo based on chapter 8 in K. Binmore., Does Game Theory Work? The Bargaining Challenge, MIT Press, Boston, 2006.

[5] Bolton G. and A. Ockenfels, 2000, ERC: A Theory of Equity, Reciprocity, and Competition, American Economic Review 90, 166–193.

[6] Carpenter J., 2004, Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods, Games and Economic Behavior, forthcoming.

[7] Ellison G., 2000, Basins of Attraction, Long Run Stochastic Stability, and the Speed of Step-by-Step Evolution, Review of Economic Studies 67, 17–45.

[8] Fehr E. and S. Gächter, 2000a, Cooperation and Punishment in Public Goods Experiment, American Economic Review 90, 980–994.

[9] Fehr E. and S. Gächter, 2000b, Fairness and Retaliation: The Economics of Reciprocity, Journal of Economic Perspectives 14, 159–181.

[10] Fehr E. and S. Gächter, 2002, Altruistic Punishment in Humans, Nature 415, 137–140.

[11] Frank R.H., 1988, Passions within Reasons: the Strategic Role of the Emotions, (Norton, New York).

[12] Levine D.K., 1998, Modeling Altruism and Spitefulness in Experiments, Review of Economic Dynamics 1, 593–622.

[13] de Quervain D.J.F., Fischbacher U., Treyer V., Schellhammer M., Schnyder U., Buck A., Fehr E., 2004, The Neural Basis of Altruistic Punishment, Science 305, 1254–1258.

[14] Steiner J., 2006, Strong Enforcement by a Weak Authority, mimeo.

[15] Yamagishi T., 1986, The Provision of a Sanctioning System as a Public good, Journal of Personality and Social Psychology 51, 110–116.