TECHNISCHE UNIVERSITÄT
BERGAKADEMIE FREIBERG
Die Ressourcenuniversität. Seit 1765.

Norm-based Sanctioning in Social Dilemmas

Sebastian Otte – Technische Universität Bergakademie Freiberg

Scientific workshop - The Role of Social Norms in Overcoming Undersupply of Public Goods 10.10.2019

Outline of the presentation

1. Introduction

2. Norm-based Sanctioning in Social Dilemmas

3. Norm-Follower Preferences

4. Reciprocity Preferences and Tangible Sanctions (Work in Progress)

5. Concluding Remarks

- Following Bicchieri (2005) a social norm is defined as a behavioral rule **R** that induces a preference for behavioral conformity with **R**…
    - …within a majority of members from a specific group **G**.
    - …for specific situations/decision problems.
    - … conditional on **empirical expectations** (belief that the majority of people in **G** conform with **R**) and **normative expectations** (belief that the majority of other people in **G** require oneself to conform with **R** and that deviators from **R** are sanctioned).

- The specific form of behavioral conformity depends on the decision problem involved
    - Discrete (binary) decision problem → **R** prescribes one action as appropriate and discards the others (Coleman 1990) → **R** induces **perfect conformity**.
    - Continuous decision problem → **R** defines a standard of behavior that determines the appropriateness of ones action (D'Adda et al. 2019) → **R** induces **partial conformity**.

- Elster (1989) sees a central feature of social norms in the application of social sanctioning:

  - Sanctioning is carried out by individuals of a group themselfs.
  - Contrary to personal norms (individual's own conscience) and legal norms (specialized legal entities).

- For the existence of a normative expectation, it is necsscary for social sanctioning to take place (or to have taken place in the past).

- Understanding the motives and different ways through which individuals carry out social sanctions is a central task for the study of social norms.

- Examples for Social Norms in Social Dilemmas:

  - **Norms of Cooperation** in Prisoner's Dilemmas which prescribe cooperation and condemn defection.
  - **Contribution Norms** in public good games that encourage contribution and define a standard for appropriate behavior (average behaviour or fixed standard).
  - **Conformity Norms** about contributive effort for public good contribution.
  - **Equity and Fairness Norms** that focus on the outcome of a game.

- In a conventional game theoretic setting (conventional preferences, one-shot games) a threat of norm-based sanctioning is not credible.

- It is here where two main theories of norm-enforcement in economics deviate in their assumptions.

2. Norm-based Sanctioning in Social Dilemmas – Theory of Repeated Games

- First theory is based on game situations changing in an indefinetly repeated version.

- If individuals are patient enough, every attainable and individual rational outcome in an indefinetly repeated prisoners dilemma game can be enforced through the use of trigger strategies by sufficiently patient individuals (Rubinstein 1979; Friedman 1971).

- Holds also in a game with limited repetitions if the last round is unclear (Axelrod 1984).

- Norm enforcement through trigger strategies of reciprocal nature:
    - GRIM-strategy: cooperate until the other/ one of the others defect, then defect yourself forever.
    - Tit-for-tat-strategy: start with cooperation, then copy the action of the other player(s) from the previous round.

2. Norm-based Sanctioning in Social Dilemmas – Theory of Repeated Games

- This approach of norm enforcement has been applied in cartel theory (e.g. Green and Porter 1984) and international environmental agreements (e.g. Barrett 1994).

- Also used by Axelrod (1984) for explaining the evolution of cooperation in the trenches of the first world war with an evolutionary approach.

- Trigger strategies based on a behavioral rule prescribing cooperation.

- An equilibrium consisting of trigger strategies is built on normative expectations: an indiviudal considering to defect expects sanctions by the other players.

- It also requires individuals to expect others to play trigger strategies. Everyone playing „always Defection" remains a viable equilibrium as well if backed by corresponding expectations (Skryms 2004).

2. Norm-based Sanctioning in Social Dilemmas – Esteem Theory of Norms

- Other approach formulated by several authors (e.g. McAdams 1995 or Pettit 1990): Indivduals care about what other people think of them.

- Their social esteem is influenced through transmission of social approval or disapproval by others as reactions to their own actions.

- Central assumption: transmission of approval happens free of cost as an emotional reaction and is based on internalized normative principles of the sanctioners.

- Individuals conform to social norms to avoid disapproval or to gain approval for their action

2. Norm-based Sanctioning in Social Dilemmas – Esteem Theory of Norms

- Has been applied in Labour Economics (e.g., Akerlof 1980; Lindbeck et al. 1999) and for some models of voluntary public good provision (e.g. Holländer 1990; Cowen 2002; Brennan and Brooks 2004).

- Cowen (2002) showed that through considering social esteem, the problem of underprovision might be aggravated in case of a contribution norm.

- Brennan and Brooks (2004) emphasized that this result depends on the nature of the behavioral standard that the generation of social esteem depends upon.

- Several possibilities:
  - A fixed normative standard tends to cause either underprovision (if it is attainable) or an overprovision (if it is unattainable).
  - A standard based on average behavior can lead to efficient provision.

2. Norm-based Sanctioning in Social Dilemmas – Esteem Theory of Norms

- Both approaches represent a good starting point for a theory of norm enforcement, but paint an incomplete picture.

- Explanatory gaps of repeated game theory:
    - only explains existence of beneficial, efficiency increasing social norms.
    - Experimental literature indicates that individuals deliver costly sanctions even in one-shot situations (Fehr and Gächter, 2002).

- Explanatory gaps of esteem theory of norms:
    - Not all individuals may care for social esteem, with some only conforming facing tangible punishment.
    - Models focus on the receivers of social esteem, not the producers.
    - Individuals deliver also costly punishments in social dilemmas.

3. Norm-Follower Preferences

- Based on Brennan et al (2013) and Bicchieri (2005) one can distuinguish two different motives involved with conforming to social norms: norm-conformists and norm-followers.

- Norm-conformists: Individuals do not care for a norm inherently but conform to avoid social sanctions or to get social rewards

- Norm-followers: Individuals care inherently for a norm and evaluate both their own and/or the norm-adherence of their peers.

- If Norm-followers witness others not conforming with the behavioral rule that they have internalized, it impacts their utility negatively.

- An example for norm-follower preferences in social dilemmas are preferences for reciprocity as defined by Rabin (1993) and Dufwenberg and Kirchsteiger (2004).

- An Individual i cares for the material outcome of a social dilemma π_i and a psychological payoff R_i that reflects reciprocity concerns. These are depicted in so-called kindness functions and enter the utility function with a weight $\alpha_i > 0$.

$$U_i = \pi_i + \alpha_i R_i$$

- Kindness functions usually bilateral relationships between a player i and j , with f_{ij} reprensenting how nice individual i expects to be to j.

- Kindness functions are dependend on i's action a_i, i's expectation about j's action b_j (first-order belief), i's expectation about what j expects that i will do c_i (second-order belief) and equitable payoffs π_i^e , π_j^e .

3. Norm-Follower Preferences – Concern for Reciprocity

$$U_i = \pi_i(a_i, b_j) + \alpha_i R_i(a_i, b_j, c_i)$$

$$R_i = f_{ij}(a_i, b_j) * f_{ji}(b_j, c_i)$$

$$f_{ij} = \frac{\pi_j(a_i, b_j) - \pi_j^e(a_i, b_j)}{\pi_j^{max}(b_j) - \pi_j^{min}(b_j)}, f_{ji} = \frac{\pi_i(c_i, b_j) - \pi_i^e(a_j, b_i)}{\pi_i^{max}(c_i) - \pi_i^{min}(c_i)}$$

$$\pi_j^e(a_i, b_j) = \frac{\pi_j^{max}(b_j) + \pi_j^{min}(b_j)}{2}$$

- $\pi_j^{max}(b_j)$: maximal possible payoff that i could provide for j, given b_j

- $\pi_j^{min}(b_j)$: smallest possible payoff that i could provide for j, given b_j

- Two players 1,2 with reciprocity preferences can contribute to a public good $G = g_1 + g_2$, with $g_i \in \{0,1\}$, at a private cost $c > 0$.

- Each unit of public good provided yields a benefit of $b < c$. It is assumed that $2b > c$.

- Without reciprocity prefences a typical Prisoners dilemma situation with Nash-equilibrium in $(0,0)$.

1,2	$g_2 = 1$	$g_2 = 0$
$g_1 = 1$	$2b - c, 2b - c$	$b - c, b$
$g_1 = 0$	$b, b - c$	$0,0$

- If both players have reciprocity preferences, the prisoners dilemma transforms to a game of coordination, if fairness preferences are sufficiently strong.

1,2	$g_2 = 1$	$g_2 = 0$
$g_1 = 1$	$2b - c + 0{,}25\alpha_1,$ $2b - c + 0{,}25\alpha_2$	$b - c - 0{,}25\alpha_1,$ $b - 0{,}25\alpha_2$
$g_1 = 0$	$b - 0{,}25\alpha_1, b - c$ $- 0{,}25\alpha_2$	$0{,}25\alpha_1, 0{,}25\alpha_2$

- (1,1) is an equilibrium for $\alpha_i > 2(c - b)$, while (0,0) is an equilibrium for every $\alpha_i > 0$.

- Players with reciprocity preferences could be implemented into esteem-theory models. They could fill the role as „producers" of social approval/disapproval.

- This needs more specified assumptions about the circumstances through which social approval is produced:

  - Approval by i coincides with the psychological effect of j's action on i → f_{ji} (argument inspired by writers of social exchange theory, like e.g. Blau 1964)
  - Approval is generated as an automatic reaction to j's action.
  - Effectiveness of approval from i depends on whether he contributes himself or not (Rege 2004).

- Player 1 with reciprocity preferences and Player 2 valuing its social esteem.

1,2	$g_2 = 1$	$g_2 = 0$
$g_1 = 1$	$2b - c + 0{,}25\alpha_1,$ $2b - c + 0{,}5\alpha_1$	$b - c - 0{,}25\alpha_1,$ $b - 0{,}5\alpha_1$
$g_1 = 0$	$b - 0{,}25\alpha_1, b - c$	$0{,}25\alpha_1, 0$

- (1,1) is an equilibrium for $\alpha_1 > 2(c - b)$, while (0,0) is an equilibrium for every $\alpha_i > 0$.

- Small model inspired by the norms game from Axelrod (1986)

- 2 Players (1,2) and two stages: First stage is a prisoners dilemma game as before. The second is a punishment stage in which each individual can choose to punish the other player, depicted by $d_i \in \{0,1\}$ through spending of $\eta > 0$, causing a tangible damage of $\gamma > 0$ for the other players material payoff.

- Two Versions:
    - 1. Player 1 homo reciprocans, Player 2 homo oeconomicus
    - 2. Both players are homo reciproans

- Sequential reciprocity by Dufwenberg and Kirchsteiger (2004): expectation about kindness is updated after the first stage has taken place.

- Utility function of Player 1 $U_1(g_1, g_2, d_1, d_2)$

- Version 1, four different subgames (g_1, g_2) to consider (1,1), (1,0), (0,1) and (0,0)

- We assume that player 2 can never credibly threat to sanction so we disregard the possibility of sanctioning by 2 here (will be discussed later).

- Utility of Player 1:

$$U_1(1,1,d_1) = 2b - c - d_1\eta + \alpha_1 \left(\frac{\frac{1}{2}b + \left(\frac{1}{2} - d_1\right)\gamma}{(b + \gamma)} \right) * \frac{1}{2}$$

$$U_1(0,1,d_1) = b - d_1\eta + \alpha_1 \left(\frac{-\frac{1}{2}b + \left(\frac{1}{2} - d_1\right)\gamma}{(b + \gamma)} \right) * \frac{1}{2}$$

$$U_1(1,0,d_1) = b - c - p_R\eta + \alpha \left(\frac{\frac{b}{2} + \left(\frac{1}{2} - d_1\right)\gamma}{b + \gamma} \right) * -\frac{1}{2}$$

$$U_1(0,0,d_1) = -d_1\eta + \alpha \left(\frac{-\frac{1}{2}b + \left(\frac{1}{2} - d_1\right)\gamma}{b + \gamma} \right) * -\frac{1}{2}$$

- Inducing punishment makes no sense in (1,1) and (0,1):

$$U_1(1,1,1) > U_1(1,1,0)$$

$$U_1(0,1,1) > U_1(0,1,0)$$

$$\alpha_1 < -2\eta \frac{b+\gamma}{\gamma}$$

$$\alpha_1 < -2\eta \frac{b+\gamma}{\gamma}$$

- Inducing punishment makes sense in (1,0) and (0,0):

$$U_1(1,0,1) > U_1(1,0,0)$$

$$U_1(0,0,1) > U_1(0,0,0)$$

$$\alpha_1 > 2\eta \frac{b+\gamma}{\gamma}$$

$$\alpha_1 > 2\eta \frac{b+\gamma}{\gamma}$$

- If it is cost-effective and reciprocity considerations are strong enough

4. Reciprocity Preferences and Tangible Sanctions

- Given this is the case, player 2 cooperates in stage 1 if

$$\pi_2(g_1, 1) > \pi_2(g_1, 0)$$

$$b(g_1 + 1) - c > bg_1 - \gamma$$
$$\gamma > c - b$$

- Punishment needs to offset the net benefit of not providing the public good

- If that is given, Player 1 cooperates if

$$U_1(1,1,0) > U_1(0,1,0)$$

$$\alpha_1 > 2(c - b)\,\frac{b + \gamma}{b}$$

- Preferences for reciprocity have to be strong enough.

- In case that reciprocity preferences are strong enough and punishment is effective, the unique sequential reciprocity equilibrium can be described in the following way:

    - Player 1 chooses $g_1 = 1$ in stage 1, while threatning to punish player 2 in stage 2 if he acts non-cooperatively and not punishing 2 if he acts cooperatively.

    - This prospect gives Player 2 an incentive to choose $g_2 = 1$. As conventional player, he will never punish in stage 2.

- In this case player 1 sucessfully enforces a social norm of cooperation that is incoporated within its reciprocity preferences.

- Other equilibria possible when $c - b > \gamma > b$ and reciprocity preferences are still strong (Full Defection with punishment) or when sanctioning is ineffective and reciprocity preferences are weak (full defection without punishment).

- Case more complicated when two players with reciprocity preferences meet each other. Both player are eligible for punishing the other in stage 2.

- Players are assumed to be symmetric, Utility of Player 1 under different histories of the game:

$$U_1(1,1,d_1,d_2) = 2b - c - d_2\gamma - d_1\eta + \alpha_1 \left(\frac{\frac{1}{2}b + \left(\frac{1}{2} - d_1\right)\gamma}{b + \gamma} * \frac{\frac{1}{2}b + \left(\frac{1}{2} - d_2\right)\gamma}{b + \gamma} \right)$$

$$U_1(0,1,d_1,d_2) = b - d_2\gamma - d_1\eta + \alpha_1 \left(\frac{-\frac{1}{2}b + \left(\frac{1}{2} - d_1\right)\gamma}{b + \gamma} \right) \left(\frac{\frac{1}{2}b + \left(\frac{1}{2} - d_2\right)\gamma}{b + \gamma} \right)$$

$$U_1(1,0,d_1,d_2) = b - c - d_2\gamma - d_1\eta + \alpha_1 \left(\frac{\frac{1}{2}b + \left(\frac{1}{2} - d_1\right)\gamma}{b + \gamma} \right) * \left(\frac{-\frac{1}{2}b + \left(\frac{1}{2} - d_2\right)\gamma}{b + \gamma} \right)$$

$$U_1(0,0,d_1,d_2) = -d_1\eta - d_2\gamma + \alpha_1 \left(\frac{-\frac{1}{2}b + d_1\gamma}{b + \gamma} \right) \left(\frac{-\frac{1}{2}b + d_2\gamma}{b + \gamma} \right)$$

- Punishment choice in different subgames depends on expectation about punishment form other player and relation of γ and b.

- Player one for $\gamma < b$

- No Punishment in History (1,1) and (0,1):

$$U_1(1,1,1,d_2) > U_1(1,1,0,d_2)$$

$$\alpha_1 < -\eta \frac{b+\gamma}{\gamma} \frac{b+\gamma}{\frac{1}{2}b + \left(\frac{1}{2} - d_2\right)\gamma}$$

$$U_1(0,1,1,d_2) > U_1(0,1,0,d_2)$$

$$\alpha_1 < -\eta \frac{b+\gamma}{\gamma} \frac{b+\gamma}{\frac{1}{2}b + \left(\frac{1}{2} - d_2\right)\gamma}$$

- Punishment in History (1,0) and (0,0):

$$U_1(1,1,1,d_2) > U_1(1,1,0,d_2)$$

$$\alpha_1 > -\eta \frac{b+\gamma}{\gamma} * \frac{b+\gamma}{-\frac{1}{2}b + \left(\frac{1}{2} - d_2\right)\gamma}$$

$$U_1(0,1,1,d_2) > U_1(0,1,0,d_2)$$

$$\alpha_1 > -\eta \frac{b+\gamma}{\gamma} * \frac{b+\gamma}{-\frac{1}{2}b + \left(\frac{1}{2} - d_2\right)\gamma}$$

- Within case $\gamma > b$ the expectation about d_2 is important

- Expectation $d_2 = 1$ in History (1,1) and (0,1) → punishment happens if α is strong enough:

$$U_1(1,1,1,1) > U_1(1,1,0,1)$$
$$\qquad\qquad\qquad\qquad\qquad U_1(0,1,1,1) > U_1(0,1,0,1)$$

$$\alpha > -\eta \frac{b+\gamma}{b} \frac{b+\gamma}{\frac{1}{2}b - \frac{1}{2}\gamma}$$
$$\qquad\qquad\qquad\qquad\qquad \alpha > -\eta \frac{b+\gamma}{b} \frac{b+\gamma}{\frac{1}{2}b - \frac{1}{2}\gamma}$$

- Expectation $d_2 = 1$ in History (1,0) and (0,0) → punishment also happens if α is strong enough:

$$U_1(1,0,1,1) > U_1(1,0,0,1)$$
$$\qquad\qquad\qquad\qquad\qquad U_1(0,0,1,1) > U_1(0,0,0,1)$$

$$\alpha > -\eta \frac{b+\gamma}{\gamma} \frac{b+\gamma}{-\frac{1}{2}b - \frac{1}{2}\gamma}$$
$$\qquad\qquad\qquad\qquad\qquad \alpha > -\eta \frac{b+\gamma}{\gamma} \frac{b+\gamma}{-\frac{1}{2}b - \frac{1}{2}\gamma}$$

- For $d_2 = 0$ in History (1,1) and (0,1) → no punishment happens

$U_1(1,1,1,1) > U_1(1,1,0,1)$ $\qquad\qquad\qquad$ $U_1(0,1,1,1) > U_1(0,1,0,1)$

$$\alpha_1 < -\eta \frac{b+\gamma}{\gamma} \frac{b+\gamma}{\frac{1}{2}b+\frac{1}{2}\gamma} \qquad\qquad \alpha_1 < -\eta \frac{b+\gamma}{\gamma} \frac{b+\gamma}{\frac{1}{2}b+\frac{1}{2}\gamma}$$

For $d_2 = 0$ in History (1,0) and (0,0) → no punishment happens

$U_1(1,0,1,1) > U_1(1,0,0,1)$ $\qquad\qquad\qquad$ $U_1(0,0,1,1) > U_1(0,0,0,1)$

$$\alpha_1 < -\eta \frac{b+\gamma}{\gamma} * \frac{b+\gamma}{-\frac{1}{2}b+\frac{1}{2}\gamma} \qquad\qquad \alpha_1 < -\eta \frac{b+\gamma}{\gamma} * \frac{b+\gamma}{-\frac{1}{2}b+\frac{1}{2}\gamma}$$

- For $\gamma > b$ the importance of the social dilemma game in stage 1 diminishes.

4. Reciprocity Preferences and Tangible Sanctions – two reciprocal players

- Sequential Reciprocity Equilibrium for this case is work in progress

- Conjecture based on first calculations :

  - In case of $b > \gamma$ similar equilibrium as in previous case with bothsided cooperation enforced through credible threats of punishment.

  - The case of $\gamma > b$ results in an equilibrium selection problem, where the players expectation about punishment by the other player matter.

- In first version, possibility of punishment by homo oeconomicus player might be important if $\gamma > b$. Refraining from punishment might be seen as „nice enough" for player one to refrain itself from punishment after witnessing defection.

- Sanctioning is an integral part of the concept of social norms, especially within the context of social dilemmas.

- Economic theory has already produced some viable approaches to analyse the enforcement of social norms, through repeated games and the esteem theory of norms.

- Because of some shortcomings of these theories (that do not undermine their importance) an additional factor is proposed with the introduction of norm-follower preferences.

- It was shown that enforcement of a norm of cooperation is possible, when individuals with reciprocity preferences can induce costly social sanctioning on a defector.

5. Concluding Remarks

- However, it was also shown that if individuals with reciprocity preferences meet each other in such a game, sanctioning may itself be the major component for the kindness functions of the player and the result from the social dilemma from before becomes less important.

- Sucessfull norm-enforcement might be questionable.

- Other problems may arise if the model is expanded in several ways:
    - n-players:
        - second order public-good problem of sanctioning
        - modeling of decentral punishment
        - Formulation of Reciprocity
        - …

    - Continuous Public good provision:
        - Formulation of reciprocity
        - Punishment also continous
        - …

Thank you for your attention!

Cited Literature

Akerlof, G. (1980), "A Theory of Social Custom, of Which Unemployment May be One Consequence", The Quarterly Journal of Economics, Vol. 94 No. 4, pp. 749–775.

Axelrod, R. (1984); The Evolution of Cooperation; Basic Books.

Axelrod, R. (1986), "An Evolutionary Approach to Norms", *The American Political Science Review*, Vol. 80 No. 4, pp. 1095–1111

Barrett, S. (1994), "Self-Enforcing International Environmental Agreements", *Oxford Economic Papers*, Vol. 46, pp. 878–894..

Bicchieri, Christina (2005); „ The Grammar of Society: The Nature and Dynamics of Social Norms"; Cambridge University press: Cambridge.

Blau, P. (1964), *Exchange and power in social life*, Wiley, New York.

Brennan, G., Eriksson, L.; Goodin, R..; Southwood, N.(2013); „Explaining Norms"; Oxford Univeristy Press: Oxford.

Brennan, G.; Brooks, M. (2004); „Esteem-based contributions and optimality in public goods supply"; Public Choice 130 (3-4); pp. 457-470.

Coleman, J. (1990); Foundations of Social Theory; Harvard University Press: Cambridge.

D'Adda, G., Dufwenberg, M., Passarelli, F. and Tabellini, G. (2019), *Partial Norms, CESifo Working Paper nr. 7568*, Munich.

Dufwenberg, M., Kirchsteiger, G. (2004), „ A theory of sequential reciprocity", *Games and Economic Behaviour 47, pp. 268-298.*

Elster, J (1989), "Social Norms and Economic Theory", The Journal of Economic Perspectives, Vol. 3 No. 4; pp. 99–117.

Fehr, E., Gächter, S. (2002); „Altruistic punishment in humans "; Nature 415;pp. 137-140.

Friedman, J.W. (1971), "A non-cooperative Equilibrium for Supergames", The Review of Economic Studies, Vol. 38 No. 1, pp. 1–12.

Green, E.J. and Porter, R.H. (1984), "Noncooperative Collusion under Imperfect Price Formation", *Econometrica*, Vol. 52 No. 1, pp. 87–100.

Holländer, H. (1990), "A Social Exchange Approach to Voluntary Cooperation", *The American Economic Review*, Vol. 80 No. 5, pp. 1157–1167.

Lindbeck, A., Nyberg, S. and Weibull, J.W. (1999), "Social Norms and Economic Incentives in the Welfare State", The Quarterly Journal of Economics, Vol. 114 No. 1, pp. 1–35.

McAdams, R.H. (1997), "The Origin, Development, and Regulation of Norms", *Michigan Law Review*, Vol. 96 No. 2, pp. 338–433.

Pettit, P. (1990), "Virtus Normativa: Rational Choice Perspectives", *Ethics*, Vol. 100 No. 4, pp. 725–755.

Rabin, M. (1993), "Incoporating Fairness into Game Theory and Economics", The American Economic Review, Vol. 83 No. 5, pp. 1281–1302.

Rege, M. (2004), "Social Norms and Private Provision of Public Goods", *Journal of Public Economic Theory*, Vol. 6 No. 1, pp. 65–77.

Rubenstein, A. (1979), "Equilibrium in Supergames with the Overtaking Criterion", Journal of Economic Theory, Vol. 21, pp. 1–9.

Skryms, B. (2004), "The Stag Hunt and the Evolution of Social Structure", Cambridge University Press: Cambridge.