Sanctioning Strategies and Internalization - a Game-Theoretic Approach*

Rolf Ziegler
Institute of Sociology, University of Munich

In his book on *„The Social System"* Talcott Parsons wrote: „Attachment to common values means, motivationally considered, that the actors have common 'sentiments' in support of the value patterns, which may be defined as meaning that conformity with the relevant expectations is treated as a 'good thing' relatively independently of any specific instrumental 'advantage' to be gained from such conformity, e.g. the avoidance of negative sanctions." (1951: 41) In a footnote he explains: „A sentiment thus involves the internalization of cultural patterns." In a contribution to the *International Encyclopedia of the Social Sciences* the psychologist Lawrence Kohlberg, who is well known for his work on stages of moral development, defines *internalization* as „learning to conform to rules in situations that arouse impulses to transgress and that lack surveillance and sanctions" (1968, Vol 10: 483).

We will try to conceptualize some of these aspects of the ***internalization of norms*** within a game-theoretic framework. The basic question is: **What kind of symbolic sanctioning strategies – praise and blame – will lead to internalization, i.e. conformity even under imperfect surveillance?**

The following example is used for illustrative purposes. Adam and Eve who are starting a small business agree to share the work load equally. The basic situation is a prisoner's dilemma game.

		Eve	
		C_E	D_E
Adam	C_A	R R	T S
	D_A	S T	P=0 P=0

Figure 1: Basic Prisoner's Dilemma Game (T>R>P>S)

Both will be rewarded (R) if they both cooperate. However, there is a temptation (T) to let the other become a sucker (S) who is doing (almost) all the work. If both deviate they will be

* Paper presented at the workshop on *„Rational Choice Sociology: Theory and Empirical Applications"*, Venice International University, San Servolo, November 30 till December 4, 2009

punished (P) because their revenue will be smaller. The incentive to deviate is measured by the difference T–R.

As we will not be able to deal with all problems involved let us briefly mention some of them which have to be solved in a more complete analysis.

- Our analysis is restricted to dyadic interactions with conjoint norms, i.e. norm targets and beneficiaries are identical, and to norms which are not externally enforced by a benevolent, disinterested philosopher-king. *Other dyadic and multi-actor situations* have to be analyzed.
- The role of *values* in determining the *content* of a norm, i.e. how the rule of sharing costs and benefits comes about, has to be conceptualized.
- The problem of *credible external sanctions* affecting an actor's physical well-being – e.g. exit, mutual defection, specific rewards and punishments – has to be taken into account.

We will focus on the use of *symbolic sanctions*, like praise and blame, affecting an actor's need for social approval, to enforce norms even under imperfect surveillance for the following two reasons. Firstly, the effectiveness of symbolic sanctions presupposes Eve being Adam's *„significant other"* and vice versa which is an important requirement for the process of internalization to work. Secondly, while applying rewards or punishments usually incurs non-negligible costs, the assumption of low or no costs of symbolic sanctions is more plausible. This reduces the problem of credibility of threats and promises.

Figure 2: Norm game with symbolic sanctions and imperfect surveillance

Figure 2 shows the *„norm game"* with symbolic positive (praise) and negative (blame) sanctions and varying degrees of *„surveillance"* to use Kohlberg's term. With an exogenously given probability *p* deviation will not be detected and sanctioned. With probability (1–*p*) surveillance is effective and the subgame has not one but two steps. At first Adam and Eve simultaneously choose to conform or to deviate and then they decide to praise (*P*) and/or blame (*B*) or not to use symbolic sanctions (*N*). We will not analyze this *„norm game with symbolic sanctions"* in detail as compared with the repeated PD-game nothing new has to be added except the fact that there exists *no* self-sustaining subgame-perfect Nash equilibrium except universal defection (even in indefinitely repeated games) if *p* gets sufficiently high, i.e. if surveillance vanishes.

At this point the question of internalization of norms arises: Under which circumstances will actors conform even if the probability *p* of any defection remaining undetected tends toward 1?

We will analyze *part* of the strategic form of an **internalization game** as we call it. But first we have to explain how pure strategies of players are defined which are assumed to be common knowledge.

In the first step both players have four *basic strategic alternatives* available:

CC unconditionally conforming

CD opportunistically conforming, i.e. conforming under surveillance, defecting if deviation goes undetected

DC defecting under surveillance, conforming if behavior goes undetected

DD unconditionally defecting

Figure 3 shows the payoffs in the basic strategic PD-game. Though both cooperating (CC,CC) would be Pareto-optimal, both defecting (DD,DD) is the only (Pareto-inferior) Nash equilibrium.

	CC	*CD*	*DC*	*DD*
CC	R R	(1–p)R+pT (1–p)R+pS	(1–p)T+pR (1–p)S+pR	T S
CD	(1–p)R+pS (1–p)R+pT	(1–p)R (1–p)R	(1–p)T+pS (1–p)S+pT	(1–p)T (1–p)S
DC	(1–p)S+pR (1–p)T+pR	(1–p)S+pT (1–p)T+pS	pR pR	pT pS
DD	S T	(1–p)S (1–p)T	pS pT	**0** **0**

Figure 3: Payoffs in the basic strategic PD-game (p = probability behavior being undetected)

In the second step the actors may use nine *ways of sanctioning* observed and unobserved (but inferred) behavior.[1]

[1] Combining these nine ways of actually and hypothetically reacting to one's opponent four first step strategies (*CC, CD, DC, DD*) in total there are therefore 9^4 = 6561 ways of reacting. Combining them with one's own four strategies in the first step will result in 4*6561 = 26.244 pure strategies. This would give a matrix of the strategic form with almost 689 million cells! However, if one assumes *consistent* behavior – i.e. the same behavior (either conformity or deviance) is never both praised or blamed – only seven ways of actually and hypothetically reacting to one's opponent four basic strategies remain. This would still result in 7^4 = 2401 ways of reacting. However, a closer look reveals that only 511 combinations fulfill the consistence criterion. Yet even this would give a matrix of the strategic form with $(4*511)^2$, i.e. more than 4 million cells.

Symbolic sanctioning strategies		Symbolic payoff
PP	praising both observed and unobserved behavior	$(1–p)P+G^{+}$
PB	praising observed and blaming unobserved behavior	$(1–p)(P+G^{+})–pG^{–}$
PN	praising observed and not reacting to unobserved behavior	$(1–p)(P+G^{+})$
BP	blaming observed and praising unobserved behavior	$–(1–p)(B+G^{–})+pG^{+}$
BB	blaming both observed and unobserved behavior	$–(1–p)B–G^{–}$
BN	blaming observed and not reacting to unobserved behavior	$–(1–p)(B+G^{–})$
NP	not reacting to observed and praising unobserved behavior	pG^{+}
NB	not reacting to observed and blaming unobserved behavior	$–pG^{–}$
NN	not reacting to neither observed nor unobserved behavior	0

Figure 4: symbolic sanctioning strategies and symbolic payoffs

It certainly has been noticed that we smuggled a *deus ex machina* into the internalization game. Without any comment we added „good and bad conscience“ (G^{+} and $G^{–}$) to the utility functions of Adam and Eve. We also assumed – without giving any reasons – that praise (P) and blame (B) have positive respectively negative utilities, though here we are in good company as it is seldom asked under which circumstances social (dis)approval has an effect. A fuller analysis has to answer the following questions:

(1) Why is ALTER a **significant other**, i.e. why does his (blame) praise have (dis)utility for EGO (which we consider to be the definition of a significant other)?
Answer: If both partners have developed *mutual respect*, praising (blaming) each other will have positive (negative) utility and will lead to a good (bad) conscience. However, without mutual respect they are not significant others and their praise (blame) will only be a sign of good (bad) reputation. As the saying goes: „In one ear and out the other“. Of course, reputation effects are important for controlling deviant behavior. However, they do not presuppose internalization. Especially a marriage impostor cares for his „good reputation“.

(2) Why does EGO have a good/bad conscience **if** his significant other **does** praise/blame him?

Answer: We agree with Ken Binmore that *„to respect oneself is simply to empathize with the respect one receives from those whose respect one reciprocates“* (1998: 269). This *definition* and the *assumption* that the amount of self-praise and self-blame is a positive function of the amount of praise and blame from a significant other have been the reason to include the terms „good/bad conscience" (G^+, G^-) in the utility function of a player *whenever* his opponent actually praises or blames him.

(3) Why does EGO have a good/bad conscience **if** his significant other **would** praise/blame him?

Answer: When Adam empathizes with the respect of Eve he realizes that his own strategies induced her to praise or blame him. He therefore realizes that she *would* praise or blame him *if* she *had* observed his behavior. Or as Adam Smith stated in *„The Theory of Moral Sentiments“* almost 250 years ago: „He anticipates the applause and admiration which in this case would be bestowed upon him, and he applauds and admires himself by sympathy – Adam Smith uses the term „sympathy“ in the same sense as „empathy“ is used here – with sentiments, which do not indeed actually take place, but which the ignorance of the public alone hinders from taking place...“ (1976 (1759): 116) For these reasons we added the terms of good or bad conscience in the counterfactual cases where Eve *would* have praised or blamed Adam *if* she *had* been able to observe his behavior.

We define a **sanctioning mode M = (m_{CC} , m_{CD} , m_{DC} , m_{DD})** as a combination of ways of sanctioning the four basic strategies CC, CD, DC and DD and we want to answer the following question:

Which sanctioning mode M = (m_{CC} , m_{CD} , m_{DC} , m_{DD}) guarantees, that (CCM,CCM) is always a (Pareto-superior) Nash-equilibrium, i.e.

(α) for all probabilities p (0 ≤ p ≤ 1) that the behavior is being undetected;

(β) with minimal requirements on the strength of external and internal symbolic sanctions P, B, G^+, G^-;

(γ) with consistent sanctioning behavior, i.e. the same behavior (either conformity or deviance) is never simultaneously praised or blamed;

will the actors always be unconditional conformists?

Without going into technical details we will briefly outline the proof which frequently compares the size of the sanctioning payoffs. From Figure 4 it can easily be deduced that the nine sanctioning payoffs form the following partial order.

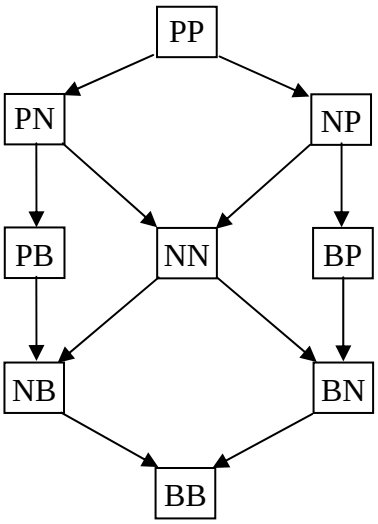

Figure 5: Partial order of sanctioning payoffs for all $0 \le p \le 1$

(1) Let us remember: In the basic game the combination (CC,CC) gives both actors the Pareto-optimal outcome R. To make the *total* payoff of an unconditional conformist CC maximal, his sanctioning payoff y_{CC} should be maximal. This will be the case if his partner chooses the sanctioning strategy PP, because – as Figures 4 and 5 shows – his sanctioning payoff then will be highest, namely $(1–p)P+G^{+}$.

(2) Now the following inequalities must be satisfied for all $0 \le p \le 1$:

(i) $(1–p)P+G^{+} \ge p(T-R) + y_{CD}$

(ii) $(1–p)P+G^{+} \ge (1–p)(T–R) + y_{DC}$

(iii) $(1–p)P+G^{+} \ge (T–R) + y_{DD}$

(3) Even if p=1, condition (iii) may be fulfilled, if y_{DD} is minimal. As Figures 4 and 5 show this will be the case if the partner chooses sanctioning strategy BB and $G^{+}+G^{-} \ge T–R$.

(4) Then – due to the consistency condition – only the following sanctioning strategies are admissible:

If ALTER chooses CD: PB, PN, NB or NN

If ALTER chooses DC: BP, BN, NP or NN

(5) Given these restrictions the minimal elements in equations (i) and (ii) are:

m_{CD} = NB with $y_{CD} = –pG^{-}$ and

m_{DC} = BN with $y_{DC} = –(1–p)(B+G^{-})$

(6) The sanctioning mode therefore is M = (PP, NB, BN, BB), i.e. deviant behavior is always blamed whether observed or not; however conforming behavior is only praised if the actor is an unconditional conformist. In the literature this mode of sanctioning is called **"sanctioning of sentiment" (Gesinnungssanktionierung).** It fulfils all three conditions (α), (β) and (γ) as long as $G^+ + G^- \geq T - R$, i.e. the strength of conscience outweighs the incentive to deviate.

There are four other modes of sanctioning which are theoretically and practically interesting:

- M = (PP, PB, BP, BB) **"sanctioning of behavior" (Verhaltenssanktionierung):** even an opportunistic conformist is praised for his observed conformity.
- M = (PP, PN, NP, NN) **"only praise"** is used as a symbolic sanction
- M = (NN, NB, BN, BB) **"only blame"** is used as a symbolic sanction
- M = (NN, NN, NN, NN) **"laissez faire"**: behavior is never sanctioned

Now the question arises: Will one of these modes of sanctioning also guarantee internalization, i.e. the strategies of two unconditional conformists define a (Pareto-superior) Nash-equilibrium even if surveillance is lacking, i.e. $p \to 1$ (assuming of course that symbolic sanctions have an effect at all, i.e. P, B, G^+ and $G^- > 0$)?

Without presenting the details of the analysis we briefly summarize the results:

- The *"laissez-faire strategy"* will never help to internalize the norm. Both players always defecting without symbolically sanctioning each other *(DDNN,DDNN)* is the only (Pareto-inferior) Nash equilibrium.
- In order to internalize the norm with the *"only praise strategy"*, the good conscience must be stronger than the incentive to deviate, $G^+ > (T - R)$; and
- In order to internalize the norm with the *"only blame strategy"*, the bad conscience must be stronger than the incentive to deviate, $G^- > (T - R)$.
- If the internal sanctions together are stronger than the incentive to deviate, $G^+ + G^- > (T - R)$ *"behavior sanctioning"* will internalize the norm as *"sentiment sanctioning"* does.
- However, if the incentive to deviate is smaller than the sum of symbolic sanctions but larger than conscience alone ($P + B + G^+ + G^- > T - R > G^+ + G^-$), *"behavior sanctioning"* will not promote internalization, but *"sentiment sanctioning"* does, as long as the probability of not being detected is sufficiently low.

$$p < \frac{(P + B + G^+ + G^-) - (T - R)}{(P + B)} \qquad \text{and} \qquad p < \frac{(P + G^+)}{(T - R) + (P - G^-)}$$

- Of course, as one would expect, if the incentive to deviate is larger than the sum of all symbolic sanctions – $(T-R) > (P+B+G^{+}+G^{-})$ – none of the modes of sanctioning will bring about internalization.

Sanctioning of sentiment turns out to be the strategy which promotes internalization under the widest conditions.[2]

It would be interesting to test some of these theoretically derived propositions with experiments.

In concluding we would like to add two *conjectures*.

Firstly, if in the medium run respect and self-respect would become equally strong the conclusion stated above could be strengthened: If social approval alone is stronger than the incentive to deviate $(P+B>T-R)$ then both players unconditionally conforming will always be a Nash equilibrium.

Secondly, if internalization is successful players will accept punishment without retaliation. This „function" of internalization – accepting sanctions even if one has the capacity to „retaliate" or even repenting and paying restitution – is one of the important prerequisites of the stability of norms besides „secondary norms" where third persons disapprove a lack of sanctioning and applaud the application of sanctions. If sanctions are not accepted by the deviant or sanctions are no longer approved and supported by third persons a norm will vanish.

[2] The results have of course to be generalized to the indefinitely repeated internalization game. However, the folk theorem guarantees that all Nash equilibria in the constituent stage game will also be Nash equilibria in the repeated game. Suitable punishment strategies may be constructed which are subgame-perfect equilibria sustaining.

References

Binmore, Ken (1994): *Playing Fair.* Cambridge, MA: The MIT Press

Binmore, Ken (1998): *Just Playing.* Cambridge, MA: The MIT Press

Kohlberg, Lawrence (1968): *Moral Development*. In: David L. Sills (Ed.): *International Encyclopedia of the Social Sciences*. New York: The Macmillan Company and The Free Press. Vol. 10: pp. 483-494

Parsons, Talcott (1951): *The Social System*. New York: The Free Press

Smith, Adam (1976): *The Theory of Moral Sentiments*. Oxford: Clarendon Press (first published 1759)

Voss, Thomas (2001): *Game-Theoretical Perspectives on the Emergence of Social Norms*. In: Michael Hechter and Karl Dieter Opp (Eds.): *Social Norms*. New York: Russel Sage Foundation, pp. 105-136

Ziegler, Rolf (2008): *Das Konzept der Internalisierung – eine spieltheoretische Analyse.* In: Andreas Diekmann, Klaus Eichner, Peter Schmidt and Thomas Voss (Eds.): *Rational Choice: Theoretische Analysen und empirische Resultate.* Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 37-53.

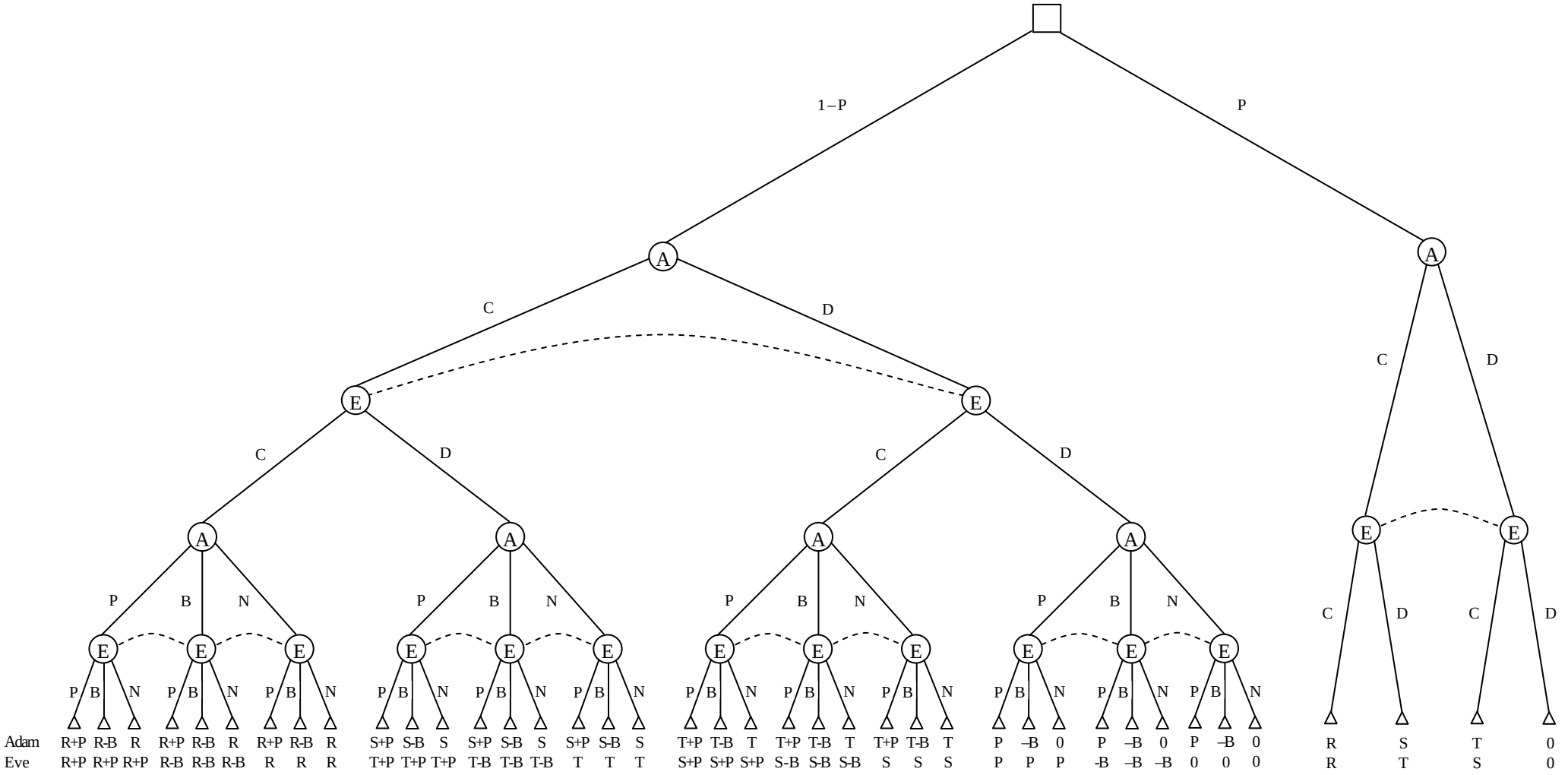

Figure 2: Norm game with symbolic sanctions and imperfect surveillance

C = conform; D = deviate; P = praise; B = blame; N = no reaction

$T > R > 0 > S$ $P > 0$ $B > 0$