Analyse & Kritik 22/2000 (© Lucius & Lucius, Stuttgart) p. 19–43

Werner Raub/Jeroen Weesie

Cooperation via Hostages*

Abstract: Conditional cooperation of selfish and rational actors is feasible in repeated encounters. We stress an important alternative for conditional cooperation: credible commitments that can be incurred via voluntary hostage posting (in the sense of pledging a bond). Hostages may facilitate cooperation in different ways. First, they reduce incentives to behave uncooperatively. Second, by offering some compensation for losses, hostages reduce the costs of suffering from uncooperative behavior of the partner. Finally, hostages may serve as signals about characteristics of the partner that are related to his opportunities and incentives to behave uncooperatively. We show that signalling hostages may have lasting effects in durable relations.

1. Introduction

The study of cooperation among 'rational egoists' goes back, at least, to the 17th and 18th century philosophers and social scientists Hobbes and Hume. Game theory, a new mathematical tool for the analysis of behavior under social interdependency developed during World War II, was applied to study the paradigmatic example of cooperation problems, namely the *Prisoner's Dilemma*. Among game theorists, it became well-known that cooperation in an infinitely repeated Prisoner's Dilemma can be maintained by conditionally cooperative strategies (this is known as the 'folk theorem'), but this finding hardly diffused to the non-specialist (Aumann 1981). In the meantime, Rapoport and his co-workers pioneered in studying how human subjects behave in one-shot and repeated Prisoner's Dilemmas (Rapoport/Chammah 1965), but this work had come to a theoretical standstill in the seventies,

* This is a thoroughly revised and extended successor of an earlier paper in Dutch that appeared in Harry Ganzeboom/Siegwart Lindenberg (eds.), Verklarende sociologie: opstellen voor Reinhard Wippler, Amsterdam 1996, 109–27. Comments of and discussions with Pawel Mlicki, Chris Snijders, and Frits Tazelaar are gratefully acknowledged. Financial support was provided by the Netherlands Organization for Scientific Research (NWO) under grant PGS 50-370. The order of authorship was established by pseudo-randomization. Our address is: Department of Sociology, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, Netherlands, email: `W.Raub@fss.uu.nl` and `J.Weesie@fss.uu.nl`.

and only social psychologists kept the spark burning with a variety of fairly ad-hoc hypotheses and studies that were not not firmly related to game theory proper. The late sixties and early seventies showed a renewed interest, especially among economists, in game theory due to major contributions to dynamic game theory (Selten 1975), and Harsanyi's discovery how games with incomplete information can be analyzed using, essentially, tools that were already there (Harsanyi 1967–8). Still, the influence of these major theoretical breakthroughs was fairly restricted. One reason is probably that the relevant literature was fairly mathematical. Also, inspiring and convincing applications were lacking.

This was the setting in which Robert Axelrod made his innovative contributions. In one respect, his work popularized results well-known in the game theoretic community. Here, Axelrod became the eloquent messenger of the good news, at least to the advocates of the free market as well as to anarchist theorists, that a utilitarian basis for cooperation may be viable on the basis of conditionally cooperative behavior. In other respects, however, Axelrod's work was much more original. Collaborating with theoretical biologists like Hamilton (1981), he introduced the theoretical developments of *evolutionary game theory* into the social sciences. Using a variety of methods—mathematical modelling, computer simulation, and the fascinating case study of the 'War of the Trenches'—he showed that a very simple form of conditional cooperation, Tit-For-Tat, has very good properties in supporting durable cooperative relations, and may even flourish in an evolutionary setting. Apart from praise, his work has, obviously, attracted lots of criticism. See for instance Hammerstein and Selten (1984) for a discussion of the stability properties of Tit-For-Tat populations and Binmore (1998) for a critical appraisal of Axelrod's computer simulations.

While it is important to appreciate that reciprocity and other forms of conditional cooperation may lead to cooperative outcomes under the right set of conditions (e.g., a sufficient duration of the relation), this does not imply that if these conditions are not met, then cooperation is not attainable. That is, one should also recognize *alternative mechanisms* that can facilitate cooperation. In this paper we analyze one such alternative mechanism. For simplicity we will focus on trust problems. These can be considered as a one-sided Prisoner's Dilemma. In a situation involving trust, the parties involved prefer an outcome where trust is placed and honored to the outcome where no trust is placed. However, placement of trust is problematic due to opportunities and incentives for abusing trust. We show how trust problems can be solved via *hostage posting* (in the sense of pledging a bond).

The remainder of this paper is organized as follows. In the next section, we summarize earlier research on trust via hostage posting. Much of that research is based on two basic assumptions. First, a trustee is assumed to have incentives as well as opportunities for abusing trust. In this specific sense, the trustee is assumed to be 'unreliable'. Second, actors are assumed to have complete information. In particular, the trustor is assumed to have information that the trustee has opportunities to abuse trust as well as information on the trustee's incentives to do so. In this situation, a hostage can generate trust by removing the trustee's incentives to abuse trust. A *binding hostage* is a hostage that does remove these incentives.

In Section 3, we analyze a more complex scenario by introducing two new and more realistic assumptions. First, we assume that actors do not have *per se* incentives and opportunities for abusing trust: there are 'unreliable' as well as 'reliable' partners. Second—and this is of course our more crucial assumption—the trustor contemplating to place trust does not know for sure whether her partner is reliable or unreliable. Hence, we consider a situation with incomplete information and, thus, bounded rationality. Posting a hostage may now not only reduce the trustee's incentives to abuse trust. A hostage may also be a *signal* indicating the trustee's reliability. We specify conditions for such a signalling effect.

Signalling is not only of interest as a substitute to conditional cooperation, but it also sheds new light on the problem of cooperation in repeated social dilemmas. Whereas Axelrod's work on the iterated Prisoner's Dilemma focuses on games with complete information, in reality information is often incomplete and asymmetric (see Rasmusen 1994 for a useful taxonomy of information problems) as players are not fully informed about the opportunities and incentives for opportunistic behavior and of the extent to which other players are far-sighted. Repetition is then important not only because it provides opportunities to reward or punish other players, but also to learn about the characteristics of the other players. Repeated games with incomplete information often have quite different properties than their complete information cousins. For instance, it is well known that in the finitely repeated Prisoner's Dilemma with complete information, there exists a unique subgame perfect equilibrium in which players defect unconditionally. The 'Gang of Four' (Kreps et al. 1982) have given an elegant example of a finitely repeated Prisoner's Dilemma in which players are uncertain whether the other player is a rational player or that he is 'programmed' to play Tit-For-Tat. They show that a 'tiny amount' of incomplete information suffices to 'rationalize' cooperation throughout many periods of the repeated game.

A well-known intuition from earlier research on trust is that *trust evolves gradually* and in a *stepwise fashion* but is *destroyed quickly*, once abused. In the fourth section of the paper, we offer a scenario where trust does indeed evolve gradually according to this intuition. We provide conditions such that posting a signalling hostage in today's interaction involving a trust problem likewise facilitates the placement of trust in future interactions.

Our analysis highlights situations with trust or—more generally—cooperation problems that are hard to solve *without* hostages or similar mechanisms mitigating and modifying *ex ante* possibilities and incentives for future opportunistic behavior. The alternative for ex ante commitments are mechanisms based on 'structural embeddedness' (Granovetter 1985; Raub/Weesie 1992) of an interaction in a sequence of interactions with the same partner or in a network of relations with third parties. Given structural embeddedness, a player can deter opportunistic behavior of the partner via threats of long-term costs associated with opportunism. These long-term costs can arise from withholding trust in future interactions. Variations of Axelrod's arguments address long-term costs due to reputation effects via social networks ('voice'; Raub/Weesie 1990; Buskens/Weesie, this volume; Buskens 1999) and termination of the relation ('exit'; see Schuessler 1989; Hirshleifer/Rasmusen 1989; Vanberg/Congleton 1992; Weesie 1992). However, the 'management' of trust problems via such mechanisms requires that future interactions are to be expected with sufficient frequency and that the short-term incentive for opportunistic behavior is sufficiently small. Our analysis refers to situations without frequently repeated interactions and with significant short-term incentives for opportunism ('golden opportunities').

2. Trust via Hostage Posting

We model simple trust relations (in the sense of Coleman 1990, chapter 5) using the trust game (TG) that was introduced by Dasgupta (1988, 59–61), Camerer/Weigelt (1988), and Kreps (1990, 100–1). Subsequently, we interpret this game as a principal-agent problem (see, e.g., Rasmusen 1994, Part II).

2.1 Modelling Trust Problems

The TG is played by two players. Player 1 is the *principal* and player 2 the *agent.* Players move sequentially. The principal moves first and must choose between placing trust in the agent and withholding trust. We denote the placement of trust by C_1 (with C indicating 'cooperation') and withholding trust by D_1 (with D indicating 'defection'). The game ends if D_1 is chosen. If trust is placed, the agent moves. The agent may honor or abuse trust. We denote honoring trust by C_2 and abusing trust by D_2. The game ends after the agent's move.

The reason why TG represents a trust *problem* becomes obvious by considering the players' preferences over the three possible outcomes of the game. According to Coleman (1990, 98–9), a typical feature of trust relations is the *risk* associated with the placement of trust. Compared to withholding trust, the principal is better off if trust is placed and honored but worse off if trust is placed and abused. Hence, the principal prefers outcome (C_1, C_2) where trust is placed and honored to outcome (D_1) where she withholds trust, while the latter outcome is preferred by the principal to (C_1, D_2) where trust is placed and abused. On the other hand, the agent has—frequently at least—incentives to abuse trust, while—just as for the principal—the situation where trust is placed and honored is more attractive for the agent than the situation where the principal withholds trust. Hence, we assume that the agent prefers (C_1, D_2) to (C_1, C_2), while preferring (C_1, C_2) to (D_1). We represent preferences of the players by (cardinal) utility functions U_i $(i = 1, 2)$. Equation 1 summarizes our assumptions on the players' preferences and introduces some additional notation:

$$\begin{aligned} U_1(C_1, C_2) = R_1 > U_1(D_1) = P_1 > U_1(C_1, D_2) = S_1, \\ U_2(C_1, D_2) = T_2 > U_2(C_1, C_1) = R_2 > U_2(D_1) = P_2. \end{aligned} \tag{1}$$

We assume that players know the structure of the game, know that the other player knows the structure of the game, and so forth. Thus, we assume that the structure of TG is 'common knowledge' (Rasmusen 1994, 44). We furthermore assume that TG is played noncooperatively: players are unable to make enforceable agreements or enforceable one-sided commitments except agreements and commitments explicitly modelled as moves in the structure (i.e., in the extensive form) of the game.[1] We use the assumptions that the structure of the game is common knowledge and that the game is played noncooperatively for all games throughout this paper.

Given our assumptions, the TG has a unique and subgame perfect Nash equilibrium (in the sense of Selten 1965: in each situation that may emerge during the game, each player's strategy maximizes his utility, given the strategy of the other player) such that the principal withholds trust (D_1) while the agent would abuse the placement of trust (D_2). This equilibrium is Pareto-suboptimal because both players would be better off had trust been placed and honored. However, due to the sequential nature of the game and the one-sided incentives and opportunities for the agent to abuse trust, placing and honoring trust is inconsistent with individually rational (equilibrium) behavior. Individually rational behavior hence implies a 'collectively irrational' (Rapoport 1974) outcome. Given the principal's and the agent's preferences, abusing trust is a typical case of 'opportunism' à la Williamson (1985, 47).

[1] Of course, this assumption is introduced because we wish to specify conditions such that rational players will be prepared to make these agreements and commitments and to specify conditions such that these agreements and commitments will stabilize the placement and honoring of trust. Thus, we employ the Nash program (Nash 1951) of explicitly modelling bargaining, communication, and all other kinds of pre-play behavior as moves in the extensive form of an extended noncooperative game, and to derive all kinds of cooperative behavior as a (kind of) Nash equilibrium of that extended game.

The principal protects herself against the agent's opportunism by withholding trust.[2]

REMARK. Social-psychologists, economists, and sociologists have conducted an enormous number of experiments with Prisoner's Dilemmas and Trust Games. A (surprisingly small) minority of these experiments involve one-shot (unrepeated) versions. While the theory predicts 'defection' (Prisoner's Dilemma) and 'not placing/abusing trust' (Trust Game), a substantial proportion of human subjects cooperate (e.g., Colman 1982, chapters 7 and 8) and place/honor trust (Snijders 1996). Most of these studies seek to find out how the proportion of cooperators varies with conditions such as the payoffs, pre-play communication, etc., with hypotheses based on informal, fairly ad hoc theory at best. Theoretically, however, it remains quite unsatisfactory that positive cooperation rates are not explained. Following Kelley and Thibaut (1978), a theory may be constructed that distinguishes between the 'objective outcomes' (e.g., money) and the 'effective outcomes' (utility) via a transformation so that utilities reflect own outcomes as well as distributional concerns, such as envy, altruism, egalitarianism, etc. Snijders (1996) provides an example of such a theoretical approach by showing that his data on Trust Games can be described parsimonuously by assuming that trustors feel guilty if they abuse trust, with guilt assumed to be proportional to a personal orientation parameter and the loss of the trustor. Moreover, subjects in the role of trustor appear to base their assessment of the guilt parameter of the trustee on their own guilt parameter ('false consensus effect'). More precisely, the behavior of the trustor depends on her own risk $(\frac{P_1-S_1}{R_1-S_1})$ and on the temptation $(\frac{T_2-R_2}{R_2-S_1})$ of the trustee, while the behavior of the trustee depends only on his temptation. For a similar analysis for 'simultaneous' 2 × 2 games, see Weesie (1994).

As an *example* for a trust game, consider a principal-agent problem from the labor market for professionals. The principal is a law firm and the agent is a newly hired lawyer, say, a specialist in 'law and information technology'. The law firm has not yet been active in this emerging field but realizes that it will become important in the future. The agent proposes to dedicate significant

[2] The trust game can be seen as a one-sided version of the Prisoner's Dilemma. This is reflected in our notation for moves and payoffs which is based on standard notation for the Prisoner's Dilemma. Note, however, that with respect to player 1 the analogy with the Prisoner's Dilemma is misleading in one respect. When we refer to 'withholding trust' by player 1 as 'defection', the analogy with the Prisoner's Dilemma is exclusively that 'withholding trust' is (part of) a Pareto-suboptimal outcome. In the Prisoner's Dilemma, each player's defection can be seen as a form of opportunistic behavior. This is true in the trust game only for player 2 (the agent). 'Defection' (withholding trust) by player 1 (the principal) in the trust game cannot be seen as opportunism but is protection against opportunistic behavior of the partner.

resources of the law firm to the exploitation of activities in this field. In such a situation, the law firm must choose between trusting the agent (C_1) or withholding trust (D_1) by making or refusing to make investments that are largely relationship-specific. These investments may include additional training and schooling of the newly hired agent, adapting the firm's internal organization to his expertise, supplying the agent with assistants and additional staff, etc. Much of these investments have to be depreciated should the agent decide to quit.

If the law firm decides to place trust by making investments, the lawyer must choose between honoring trust (C_2) through a durable relationship with the law firm as an associate or as a partner and abusing trust (D_2) by using his new appointment as a stepping stone for a more attractive job offer of another firm. Assuming that the lawyer is one of a few highly qualified specialists in a newly emerging and important field, it seems a real possibility that he has—or will have—such exit options via outside offers.

Considering the preferences of the law firm, it seems likely that the situation with placement of trust that is honored by the agent (C_1, C_2) is preferred to withholding trust (D_1). Otherwise, the law firm would not even consider making the investments mentioned. On the other hand, an outcome such that trust is placed and abused (C_1, D_2) is even worse for the law firm than withholding trust, at least if the required investments are sufficiently large and relationship-specific. The newly hired lawyer, conversely, would profit from the firm's investments if he stays with the firm for a sufficient period of time. Hence, honoring trust (C_1, C_2) is a more attractive outcome also for the agent than the outcome with the principal withholding trust (D_1). However, a typical implication of, e.g., additional training and schooling offered by the principal will be that the agent's market value for other firms increases so that the agent's most preferred outcome might indeed be to abuse trust (C_1, D_2).

2.2 Hostages as a Solution of Trust Problems

In a trust relation like the one we have modelled up to now, the pricipal as well as the agent are facing a problem. *Both* suffer if trust is withheld. We now proceed to specifying conditions such that a rational agent and a rational principal are able to solve their trust problem. Hence, we consider generating and stabilizing trust as our *explanandum.*[3]

Conditions for generating and stabilizing trust are related to the embeddedness of a trust relation in a social context (Raub/Weesie 1992). As argued

[3] See Craswell 1993 for a useful discussion of the difference between an analysis that focuses on the results of trust—so that trust is part of the explanans—and an analysis like in our paper where 'trust' is not assumed as given but is considered itself as the explanandum.

in the introduction, we focus on one specific condition for solving trust problems, namely, the availability of a hostage.[4] Hence, we contribute to institutional analysis and institutional design (see for the following Weesie/Raub 1996, 203–5). Institutions can be broadly conceived as constraints for human action that result from human action itself and structure the incentives in social relations (see, e.g., North 1990, chapter 1). We take as given and exogenous an institutional context that provides opportunities for the agent to post a hostage *ex ante*, i.e., before the principal decides to place or to withhold trust. The agent loses the hostage if he abuses trust later on. Posting a hostage constitutes a 'strategic move' (Schelling 1960): If the hostage is sufficiently valuable for the agent, the hostage modifies the agent's incentive to abuse trust. By posting a hostage, the agent can incur a '*commitment*'. The hostage can be a safeguard for the principal that the agent would honor rather than abuse trust. A context providing an *option* to post a hostage is an initial condition in the subsequent analysis. Such an opportunity can be considered as '*institutional capital*' of the players. By using this capital, e.g., by posting a hostage, players create their own private institutions or, as Coleman (1990) put it, their 'constructed social environment' that can facilitate placing and honoring trust. The private institutions themselves are endogenous in our analysis. We provide conditions such that these private institutions are self-enforcing, i.e., result from individually rational (equilibrium) behavior (see Schotter 1981 and Calvert 1992 for the distinction between institutions as constraints and institutions as (an outcome of) equilibrium behavior). Hence, we do *not* assume that an external third party *forces* an agent to incur a commitment via hostage posting. We address the 'deeper' question concerning conditions such that an agent posts a hostage *voluntarily* and without external coercion.

To specify conditions for posting a hostage and for generating and stabilizing trust via hostage posting, we introduce a hostage game HTG which is an extended version of the original TG. In HTG, the agent moves first. He chooses between posting or not posting a hostage. Subsequently, the principal is informed on the agent's move and the players play the trust game TG. Thus, the principal first decides to place or to withhold trust. If trust is placed, the agent decides to honor or abuse trust. When the principal moves, she knows whether the agent has posted a hostage. Hence, the principal can condition her own move on the hostage posting decision of the agent.

We assume that the agent loses his hostage if and only if he posts it at the beginning of the game, if the principal places trust, and if the principal's trust

[4] Schelling 1960 and Williamson 1985 highlighted the use of hostages for manipulating the outcomes of social interactions, particularly in the context of distribution problems. See Weesie/Raub 1996 and Raub/Keren 1993 for earlier theoretical and experimental work on the use of hostages for solving cooperation problems. Snijders 1996 focuses specifically on hostages as a mechanism for solving trust problems.

is finally abused by the agent. If the agent's hostage is lost, it is *not* given to the principal. Hence, we exclusively focus on the effects of hostage posting for the agent's payoffs. We neglect situations such that a hostage is given to the victim of opportunistic behavior and, thus, also has a *compensating* effect (see Raub/Keren 1993, Weesie/Raub 1996, and Snijders 1996 for the theoretical and empirical impact of such a compensating effect). The payoffs for the principal in HTG do not depend on the hostage posting decision of the agent but exclusively on behavior in the subsequent trust game TG itself. The payoffs for the agent in HTG do of course depend on the agent's hostage posting decision. We assume that the hostage has value $K > 0$ for the agent. We also assume that hostage posting is associated with (transaction) costs $\tau \geq 0$ for the agent. These costs arise if the agent decides to post a hostage. Hence, these costs are not only due if the agent loses his hostage (in this case, one could consider these costs simply as an ingredient of the value K of the hostage) but also if trust is placed and honored after a hostage has been posted. We assume that the agent's utility at the end of HTG is additive in his payoff at the end of the corresponding trust game TG, in transaction costs τ, and the value K of a lost hostage. For example, the agent's utility is $T_2 - K - \tau$ if he posts a hostage and subsequently abuses trust that has been placed by the principal, $R_2 - \tau$ if he posts a hostage and subsequently honors trust that has been placed by the principal, and P_2 if the agent does not post a hostage and the principal withholds trust.

Our first theorem offers a sufficient condition for hostage posting, placing trust, and honoring trust by a rational agent and a rational principal.

Theorem 1 *Trust via hostage posting (Weesie/Raub 1996). HTG has a subgame perfect equilibrium such that the agent posts a hostage, the principal places trust, and the agent honors trust if*

$$K > T_2 - R_2 \quad \textit{and} \quad \tau < R_2 - P_2. \tag{2}$$

This simple theorem confirms our intuition that it can be individually rational for the agent to voluntarily post a hostage and that the hostage can be a sufficient safeguard for the principal to place trust. This is the case if the hostage is sufficiently valuable for the agent so that after hostage posting he has incentives to honor trust if trust is placed. Hence, the hostage has to be 'binding'. The critical value for K are the costs $T_2 - R_2$ for the agent of honoring trust. Moreover, the transaction costs associated with using the hostage mechanism have to be sufficiently low. The critical value of τ is the efficiency gain $R_2 - P_2$ associated with placing and honoring trust compared to withholding trust. Note that the equilibrium where posting a hostage induces the principal to place trust and where the agent honors trust is a Pareto-improvement compared to the 'no trust'-equilibrium of the original trust game TG. However, while this equilibrium is a Pareto-improvement, it is Pareto-optimal only if no transaction costs are associated with hostage posting ($\tau = 0$).

Note also that the equilibrium strategies that support hostage posting, placement of trust, and honoring trust according to Theorem 1 are in one respect similar to conditional cooperation à la Axelrod in a repeated game. Namely, the equilibrium strategies are 'reactive' in the sense that both players condition own behavior on the behavior of the partner. The equilibrium strategy of the principal makes the placement of trust dependent on prior hostage posting of the agent. The principal's equilibrium strategy could be interpreted as the tacit promise to place trust after hostage posting and the tacit threat to withhold trust if the agent does not post a hostage. Conversely, the agent's equilibrium strategy implies a tacit promise not to abuse the principal's trust. Subgame perfectness of the equilibrium makes for credibility of these threats and promises.

Our analysis reveals that the meaning of 'unreliability' of the agent is ambiguous. First, 'unreliability' may mean that the agent has opportunities and incentives for abusing trust in the underlying TG because he can choose between C_2 and D_2 and because $T_2 > R_2$. In this sense, 'unreliability' refers to the ***structure of the interaction situation*** for the agent. Second, 'unreliability' may mean that the agent would actually abuse the principal's placement of trust: the agent would choose D_2. In this sense, 'unreliability' refers to actual or potential *behavior*. Theorem 1 shows that an agent who is 'unreliable' in the first sense may well be 'reliable' in the second sense because he would actually honor trust. In the following, 'unreliability' refers exclusively to the *feasibility of and to incentives for opportunistic behavior* and *not* per se to an agent actually succumbing to such a temptation.

REMARK. Various experiments have been conducted on hostage posting in Prisoner's Dilemmas and Trust Games. Raub/Keren (1993) found experimentally that subjects are more likely to post a hostage in a symmetric one-shot Prisoner's Dilemma if such behavior is part of the (Pareto-dominant) subgame perfect equilibrium to do so, and this tendency is even stronger if it also constitutes maximin behavior. Mlicki (1996) studies how the use of hostages depends on transaction costs associated with the hostage mechanism, and on extra rewards from cooperation if hostages are posted (so-called 'productive hostages', or 'relationship-specific investments'). Using a similar experimental design as Raub/Keren, Mlicki found that subjects are more likely to post a hostage the lower the transaction costs and the more productive the hostage. These effects are also found for small transaction costs and productivity bonuses that do not affect the predictions derived from a game theoretic analysis. Snijders (1996) conducted experiments on hostage posting in Trust Games. He showed that part of the effects of the hostages can be explained by the way in which hostages affect the indices *risk* and *temptation* that describe behavior well. These effects of hostages reflect the bonding and compensatory mechanisms. However, the possibility to post a hostage appeared to affect behavior stronger than can be accounted for by the modification of the payoffs for the consecutive Trust Game. Theoretically, these have to be signalling effects analyzed in the next section.

Let us return to our example: the interaction between the law firm and the newly hired specialist for 'law and information technology'. How can the agent post a hostage which is considered a sufficient safeguard by the principal for a durable employment or partnership relation with the agent and hence induces the principal to invest in the relation? A typical option for the agent is moving and acquiring real estate close to his new job. Posting such a hostage is associated with considerable transaction costs. These include financial costs for a real estate agent and a conveyancer, renovation and redecoration costs, moving costs, but also social costs from losing relations of the agent, his partner, and children that are tied to his former place of residence. Finally, the agent incurs costs by undermining his bargaining position should a new outside offer emerge. On the other hand, moving and acquiring real estate clearly constitutes a hostage: if the agent quits prematurely, he will be confronted with costs of the same type as those mentioned and this makes acceptance of a new appointment elsewhere less attractive. Employers frequently capitalize on this situation. E.g., they reimburse employees for moving costs and thus reduce the employees' transaction costs associated with posting the hostage. Likewise, the reimbursement for moving costs is often conditional on a minimum employment duration so that the value of the hostage increases for the employees.

Our example of hostage posting for stabilizing trust in a principal-agent relation reveals a number of interesting features of the hostage mechanism. Notice first that hostages can be posted 'informally'. Contractual arrangements are not necessary. A contract between employer and employee stipulating that the employee is not allowed to quit, at least not before a certain period, would not be legally binding and enforceable. Moving is an informal, non-contractual way of posting a hostage. Second, our example shows that hostages can be posted without interventions of outside third parties. There is no third party like the state, a broker, or a notary in our example that is needed to take charge of the hostage. Third, the example demonstrates that even if such a third party is not available it can be possible for the agent to post a hostage without incurring the 'expropriation hazard' (Williamson 1985, 177) that the principal refuses to return the hostage even if the agent has honored trust.

Obviously, our model of trust via hostage posting could be extended in various ways.[5] We mention but one possible complication, a variant of William-

[5] For some extensions, see Raub/Weesie 1993; Snijders 1996; Weesie/Raub 1996.

son's *expropriation risk*, which is largely neglected in our analysis. We have assumed that only the agent and not the principal is facing possibilities and incentives for opportunistic behavior. Our example shows, however, that reality might be more complex in the sense that also the employer may be tempted to abuse trust placed by the employee. A typical case is that the offer of the law firm for the specialist includes facilities for future training and schooling and, particularly, promises with respect to the future general policy and strategy of the law firm. Such promises are frequently non-contractual and not legally binding or enforceable. If the agent places trust in informal promises and commits himself by moving, he loses flexibility to react to the law firm's future deviations from the promises. In a completely different context, Becker (1991, 12–3) has mentioned that posting hostages and, more generally, incurring commitments implies a loss of flexibility in reacting to unexpected contingencies. We disregard a thorough analysis of this problem and simply assume that the expected costs of a loss of flexibility are included in the transaction costs associated with posting the hostage.

3. Incomplete Information and Hostages as Signals

Up to now, we have assumed that the agent always has opportunities and incentives for abusing trust. Thus, we have assumed that the interaction situation is such that the agent is 'unreliable'. A more realistic—but likewise more complex—assumption is that the agent faces such opportunities and incentives not with certainty but only with some positive probability. In other words, the interaction situation may be such that the agent is in fact 'reliable'. Concerning the players, we have assumed that the principal has complete information about the possible actions as well as the incentives of the agent. A more realistic—and again more complex—assumption is that the principal has *incomplete information*: she (only) knows the *probability* such that the agent has an option and an incentive for opportunistic behavior but not the agent's *actual* behavioral alternatives and incentives. According to Williamson (e.g., 1985, 46, 81), such incomplete information is a typical feature of 'bounded rationality'.

What are possible implications of such a scenario for the use of hostages as a mechanism for solving trust problems? On the one hand, a hostage can still modify the incentives for an unreliable agent, i.e., the hostage can still be binding. On the other hand, the hostage may now also have the function of a signal. Posting a hostage may signal that the agent is in fact reliable in the sense that he has no option and no incentive for opportunistic behavior.[6] In this case, hostage posting affects the principal's information about char-

[6] Schelling 1960 and Williamson 1985 suggested that various forms of uncertainty and

acteristics of the agent. Obviously, hostages that serve signalling purposes contribute to the 'definition of the situation' and to 'framing'. These are classical topics of (micro-) sociology. Recently, it has been tried to integrate these phenomena into a rational actor perspective (e.g., Lindenberg 1992; Esser 1993). However, a more rigorous model for the analysis of 'relational signals' (Lindenberg 1994, 106–8) is still lacking. One such model is provided in the following.

3.1 A Trust Game with Incomplete Information

We first consider a trust game TGI such that the agent faces a temptation for opportunistic behavior only with some positive probability and where the principal has incomplete information about the situation of the agent. We model such a scenario as a game with incomplete information (see Rasmusen 1994). We neglect hostages in this section.

In TGI the principal can meet two possible 'types' of agents. The first type is 'unreliable' in the sense that such an agent has action alternatives and incentives as in the original trust game TG: if the principal places trust, the agent may either honor (C_2) or abuse trust (D_2) and abusing trust yields a higher utility (T_2) for the agent than honoring trust (R_2). An agent of the second type is 'reliable', i.e., he has no opportunities for abusing trust: if the principal places trust in such an agent, the game ends with payoff R_i for both players.[7] A chance move of Nature at the beginning of TGI determines the type of agent playing the game. With probability π ($0 < \pi < 1$), Nature 'chooses' an unreliable agent and with probability $1 - \pi$ a reliable one. While π is assumed to be known to both players, the outcome of Nature's initial move, i.e., the agent's *actual* type, is unobservable for the principal. Hence, TGI is a game with 'incomplete information'. The agent, on the other hand, is informed on his own type so that TGI is also a game with 'asymmetric information'.

After Nature has determined the agent's type, the principal moves, just like in the original TG. She chooses between placing (C_1) and withholding trust (D_1). The game ends if the principal withholds trust. In this case, both players receive payoff P_i. If the principal places trust and the agent is of the reliable type, the game ends likewise and both players derive utility R_i. If the principal places trust and the agent's type is 'unreliable', the agent chooses between honoring (C_2) and abusing trust (D_2). If he honors trust,

incomplete information may affect the use of hostages in social interactions. Snijders 1996 provides various formal analyses.

[7] An obvious alternative scenario would be that a reliable agent has opportunities but no incentives to abuse trust. This would be the case if the agent has internalized norms and values inducing sufficient 'internal sanctions' should the agent behave opportunistically. Then, honoring trust would be associated with a higher 'net utility' than abusing trust. Such an alternative scenario leads to similar results like those presented in this paper.

both players receive payoff R_i. Should he abuse trust, the payoff for the principal is S_1 while the agent receives T_2. Afterwards, the game ends.[8]

Returning to our example of a principal-agent problem, the interpretation of TGI is rather obvious. With probability π, the newly hired specialist does have or receive an outside offer and with probability $1 - \pi$ he does not have such an exit opportunity. The probability for an outside offer depends on characteristics of the labor market as well as on characteristics of the specialist himself. We assume that all these characteristics and hence the probability π itself are well-known for the law firm. However, the law firm does not know if the labor market has or has not actually generated the 'golden opportunity' of an outside offer for the specialist.

What is individually rational behavior in the TGI? The equilibrium is of course such that an unreliable agent abuses trust because $T_2 > R_2$. The principal chooses to withhold trust if the payoff P_1 she receives after this move is larger than her expected payoff if she places trust. This is the case if

$$U_1(D_1) = P_1 > U_1(C_1) = \pi S_1 + (1 - \pi)R_1 \tag{3}$$

which can be rewritten as

$$\pi > \frac{R_1 - P_1}{R_1 - S_1}. \tag{4}$$

The trust game TGI with incomplete information thus has a Pareto-suboptimal solution and individually rational behavior implies a collectively irrational outcome if the probability π of meeting an unreliable agent is sufficiently large, e.g., if demand on the labor market for the agent's expertise or the quality of his expertise are sufficiently high. The critical value for the probability of an unreliable agent is specified in (4).[9] In the following, we assume that π fulfills (4) so that a trust *dilemma* emerges in TGI: the principal withholds trust although the situation where trust is placed and not abused would be more beneficial for the principal as well as for both types of agents (and although trust *could not* even be abused by a reliable agent).

[8] Note that a reliable agent never moves in TGI. Note also that the original trust game TG results from TGI for $\pi = 1$. In this case, the agent is always unreliable and the principal is completely informed on the agent's type.

[9] Coleman 1990, 99, provides a condition such that trust is placed by the trustor in a simple trust game like TG. Surprisingly, Coleman neglects a strategic analysis of trust problems and analyzes the trustor's decision situation as a game against Nature and not as a game against an incentive-guided partner. Note that Coleman's condition is a special case of inequality (4). Hence, our analysis provides a 'rational reconstruction' of Coleman's 'trust condition'. The results of Snijders' 1996 experiments show that—empirically—the ratio $\frac{R_1-P_1}{R_1-S_1}$ in (4), i.e., the trustor's risk, is crucial for describing and explaining behavior in trust games.

3.2 Hostages as a Solution for Trust Problems in Situations with Incomplete Information

Specifying conditions such that hostages can solve trust problems in situations with incomplete information becomes feasible after extending TGI once more with an option for the agent to post a hostage. This extension is modelled in the hostage game HTGI. Again, this is a game with incomplete and asymmetric information. Just as in TGI, Nature first determines the agent's type, where Nature's move is observed by the agent (he knows his own type), while the principal cannot observe the agent's type and is only informed on the respective probabilities for either type of agent. Subsequently, the agent may or may not post a hostage. The principal receives information on the agent's hostage posting decision, just like in the hostage game HTG. Afterwards, the principal decides to place (C_1) or to withhold trust (D_1). The game ends after D_1. Just like in the trust game TGI with incomplete information, the game ends after C_1 if the agent is of the reliable type. If the principal places trust and the agent is unreliable, the agent chooses between honoring (C_2) and abusing trust (D_2) and the game ends.

Assumptions on payoffs for the players in HTGI correspond to the assumptions used for HTG and TGI. Hence, we assume again that the agent loses his hostage if and only if it is posted at the beginning of the game and if trust is placed and abused (this implies of course that a reliable agent never loses his hostage). If the hostage is lost, it is not given to the principal: the principal's payoff does not depend on the hostage posting decision of the agent. Thus, we consider hostages that have, as we will see below, signalling and bonding characteristics, but not with compensatory properties. The payoff of the agent does depend on his hostage posting decision. The agent's payoff at the end of HTGI is again additive in his payoff at the end of the corresponding trust game, transaction costs, and the value of a lost hostage.

We assume that the value of the hostage is the same for both types of agents and equal to K. Posting a hostage is again associated with transaction costs for the agent. Again, these costs emerge always if the agent decides to post a hostage and do not depend on what happens later on in the game. The *crucial assumption* is that the *transaction costs may differ, depending on the type of agent.*[10] By allowing for differences in transaction costs, we can use the fundamental insight of signalling theory (Spence 1974) that differences

[10] Similar results would be obtained if the transaction costs are the same for the two types of agents, while losing the hostage is more costly for the unreliable actor than for the reliable actor, or if the reliable agent has a higher 'guilt parameter' (see the remark in Section 2.1). Without any differences in the consequences of posting a hostage, signalling effects are not possible, as the unreliable agent may always mimic the reliable agent if posting a hostage convinces the principal of the type of agent, thereby invalidating the beliefs of the principal.

in signalling costs can have far-reaching ramifications for signalling behavior and the credibility of signals. We denote the costs of hostage posting for an unreliable agent with τ_u. Cost of hostage posting for a reliable agent are denoted with τ_r. The interesting question is whether differences in these costs can imply that posting a hostage by the agent allows for conclusions of the principal about the agent's type. In particular, we will consider whether such 'definitions of the situation' or 'framing-effects' can emerge if the transaction costs for a reliable agent are smaller than those for an unreliable agent, i.e.,

$$\tau_r < \tau_u. \tag{5}$$

In the context of our example, such a difference is rather likely. An important ingredient of the transaction costs involved in posting a hostage by acquiring real estate results from undermining one's bargaining position vis-à-vis an alternative employer offering a new appointment. These transaction costs emerge by definition for an agent who is unreliable while an agent who is reliable does not have to incur these costs.

In a scenario like the one modelled via HTGI, hostages can serve two different purposes (see Mlicki/Snijders 1995). First, just like in HTG, posting a hostage ex ante may modify the preferences of an (unreliable) agent such that honoring trust becomes individually rational ex post. Hence, the hostage can still be binding. Moreover, however, posting a hosting may now indicate the agent's type, i.e., the hostage may serve as a signal. Of course, the hostage can serve as a signal only if a rational and reliable agent would post it, while a rational and unreliable agent would not.

Hostage posting, and subsequent placement of trust by the principal is individually rational in HTGI if these moves are supported by a 'Bayesian' equilibrium (roughly, an equilibrium such that the principal uses her observations of the agent's behavior for updating her beliefs about the agent's type in a rational way, i.e., according to Bayes' rule). Two kinds of equilibria are pertinent. First, we consider an equilibrium such that both types of agents post a hostage and the principal subsequently places trust. This is a '*pooling equilibrium*': both types of agents behave in the same way as far as hostage posting is concerned. In this case, the hostage is binding but does not signal the agent's type. Of course, we are more interested in another kind of equilibrium such that only a reliable agent posts the hostage and the principal places trust only if a hostage has been posted. This is a '*separating equilibrium*': the reliable agent's behavior differs from the behavior of the unreliable agent. The following theorem provides sufficient conditions for a pooling as well as a separating equilibrium.

Theorem 2 *Binding and signalling hostages. HTGI has a pooling equilibrium such that both types of agents (a) post a hostage, (b) the principal places trust, and (c) an unreliable agent honors trust if*

$$K > T_2 - R_2 \quad \textit{and} \quad \max(\tau_r, \tau_u) < R_2 - P_2. \tag{6}$$

Moreover, HTGI has a separating equilibrium such that (a) a reliable agent posts a hostage, while an unreliable agent chooses not to post a hostage and (b) the principal places trust after a hostage has been posted and withholds trust if no hostage has been posted if

$$K > T_2 - R_2 \quad \textit{and} \quad \tau_r < R_2 - P_2 < \tau_u \tag{7}$$

or

$$K < T_2 - R_2, \quad \tau_r < R_2 - P_2, \quad \textit{and} \quad T_2 - P_2 - K < \tau_u. \tag{8}$$

Both equilibria specified in Theorem 2 support strategies of the principal such that trust is placed only if a hostage has been posted. This follows immediately from condition (4). In the pooling equilibrium in Theorem 2, the hostage of an unreliable agent is sufficiently valuable and, thus, binding: the unreliable agent has incentives to honor trust if he posts a hostage and the principal places trust. Moreover, transaction costs are sufficiently small for both types of agents. Because a hostage is posted by both types of agents, hostage posting cannot signal an agent's type.

HTGI has a separating equilibrium if either (7) or (8) are fulfilled. Given (7), a hostage is again binding. Should an unreliable agent post the hostage, he would also honor trust. However, due to (7), the costs of hostage posting for an unreliable agent are such that the situation where he honors trust after having posted a hostage is less attractive for him than the situation where no hostage has been posted and the principal withholds trust. Conversely, according to (7), the costs of hostage posting are sufficiently small for a reliable agent so that his payoff after he has posted a hostage and trust has been placed by the principal exceeds the reliable agent's payoff if no hostage has been posted and trust is withheld by the principal. Hence, posting a hostage now becomes a (reliable) signal of an agent's type and it becomes rational for a reliable agent to signal his type via hostage posting. We see that sufficient differences in transaction costs for the two types of agents imply a signalling function of the hostage. In terms of our example, if moving sufficiently undermines the bargaining position of the lawyer vis-à-vis an outside job offer, he will move only if he does not have an outside offer. In this case, the law firm will make relationship-specific investments only for new associates or partners without an exit-option.

Finally, consider implications of (8). From the perspective of 'hostages as reliable signals' these conditions are most revealing. Due to the first condition in (8), the hostage is no longer binding, i.e., an unreliable agent would abuse trust if trust has been placed by the principal after hostage posting of the agent. In this situation, the principal will place trust after hostage posting by the agent only if she can conclude from hostage posting that the agent is reliable. Such a conclusion, however, can indeed be derived under the conditions of our separating equilibrium. According to the second condition in (8), the costs of hostage posting for a reliable agent are small enough so that his payoff after hostage posting and placement of trust exceeds his payoff if he posts no hostage and the principal withholds trust. Conversely, due to the last condition in (8), it is not attractive for the unreliable agent to imitate the reliable agent by posting a hostage and subsequently to react opportunistically should the principal decide to place trust (of course, if such imitation would be attractive for the unreliable agent, the strategy of the principal would not be optimal). Hence, posting a hostage is again a reliable signal of the agent's type even though the hostage is not even binding. Again, it becomes rational for the reliable agent to signal his type. Notice that the signalling effect of hostage posting depends once more on differences in transaction costs.

Just like in the hostage game HTG with complete information, the strategies underlying the equlibria from Theorem 2 are 'conditional': players condition their behavior on prior behavior of the partner and use tacit and credible threats and promises. Moreover, the equilibria again constitute 'second best' solutions and Pareto-improvements compared to the 'no trust'-equilibrium in the original trust game.

4. Hostages and the Stepwise Evolution of Trust

Up to now, we have analyzed how trust can be generated in one-shot interactions. If actors entertain durable relations it seems unlikely that the 'size' of their trust problems and, hence, the resources they have to invest in order to solve these problems remain constant throughout their relation. Rather, it has been suggested that trust evolves *gradually* and in a *stepwise fashion*, and is likewise quickly destroyed if it has been abused. As Blau (1964, 94; see Dasgupta 1988, 62 and Coleman 1990, 104 for related intuitions) put it: "...exchange relations evolve in a slow process, starting with minor transactions in which little trust is required because little risk is involved ...[P]rocesses of social exchange, which may originate in pure self-interest, generate trust in social relations through their recurrent and gradually expanding character." From a theoretical perspective, gradually 'increasing' placement of trust by the principal may result from changing incentives of the actors like reduced incentives of the agent to abuse trust or reduced costs for the principal if trust is abused. However, trust may also evolve due to changing anticipations of the principal with respect to characteristics of the agent. Mlicki/Snijders (1995, 15–6) pointed out that posting a hostage with signalling properties precisely has an impact on the principal's beliefs and expectations. Posting a hostage may then have a durable effect on the principal's propensity to trust via a durable influence on the principal's information about characteristics of the agent.

To make these intuitions precise, we again extend the scenario from Section 2. We do so by introducing another hostage game, HTGII. The difference between the hostage game HTGI from Section 3 and HTGII is that a second trust game is played after the first one. The hostage only affects the agent's payoffs in the first trust game and the agent cannot post another hostage before the second trust game is played. In terms of our example, consider a situation such that the law firm must decide again on whether or not to make relationship-specific investments, e.g., by hiring additional staff for the field of 'law and information technology'.

The structure of HTGII is as follows. First, HTGI is played: Nature determines the agent's type and the principal is unable to observe the outcome of Nature's move. Afterwards, the agent chooses between posting and not posting a hostage, which decision the principal is able to observe. The principal then decides to place or to withhold trust and if trust is placed, the unreliable agent chooses between honoring and abusing trust. Subsequently, the principal receives information on her own payoff (R_1, P_1, or S_1) but not on the possible move of the agent.[11] Finally, another trust game TG is played. We assume that the possible payoffs in this TG are again R_1, P_1, or S_1 for the principal and T_2, R_2, or P_2 for the agent.[12] The payoffs of both players in HTGII are additive in their payoffs for the two trust games, possible transaction costs associated with hostage posting prior to the first trust game and the value of a possibly lost hostage after abusing trust in the first trust game.

Under such a scenario, a rational and unreliable agent would always abuse trust in the second trust game, even if he may have honored trust in the first trust game. This is due to the fact that he cannot commit himself for the second trust game. Hence, two equilibria are particularly interesting in HTGII (leaving aside the trivial case without hostage posting by the agent and a principal withholding trust in both trust games). First, we are interested in a pooling equilibrium such that both types of agents post a hostage that is not signalling, the principal places trust in the first trust game which will

[11] We wish to avoid the too simple 'solution' for our problem that the principal can infer the agent's type from observing that the agent 'is doing nothing' (does not move) after trust has been placed. For the same reason, we exclude the possibility that the principal can infer the type of an *un*reliable agent from the observation that the agent '*is* doing something', namely, honors (!) trust in the first trust game.

[12] The simplifying assumption that payoffs are the same in both trust games is not really needed for the following analysis. Assuming 'heterogeneous trust games' (see Raub/Weesie 1993) would lead to similar results.

be honored by the unreliable agent, and the principal withholds trust in the second trust game. Given this equilibrium, hostage posting has an effect for the first trust game but does not contribute to building up trust for new situations: the hostage does not have a 'transfer-effect'. Second, consider a separating equilibrium. A reliable agent posts a hostage while an unreliable agent refuses to do so, and the principal places trust in both trust games only if a hostage has been posted at the beginning. In this separating equilibrium, the hostage has an effect for *both* trust games.[13] The following theorem provides sufficient conditions for a pooling as well as a separating equilibrium.

Theorem 3 *Hostages and gradually expanding trust. HTGII has a pooling equilibrium such that (a) both types of agents post a hostage, (b) the principal places trust in the first trust game, (c) the unreliable agent honors trust in the first trust game, and (d) the principal withholds trust in the second trust game if (6) is fulfilled.*

Moreover, HTGII has a separating equilibrium such that (a) only a reliable agent post a hostage and (b) the principal places trust after hostage posting in the first and in the second trust game if

$$\tau_r < R_2 - P_2 < \tau_u - (T_2 - P_2). \tag{9}$$

The pooling equilibrium in Theorem 3 has the same interpretation as the pooling equilibrium in Theorem 2. Posting a hostage has no signalling effect in this equilibrium and due to (4) the probability of meeting an unreliable agent is large enough so that trust is withheld in the second trust game.

The separating equilibrium in Theorem 3 reveals that hostages may not only contribute to generating and stabilizing trust in one-shot transactions but may also induce a gradual expansion of trust. The equilibrium is somewhat similar to the separating equilibrium from Theorem 2. The important condition with respect to transaction costs of the unreliable agent is once more that these are large enough so that he has no incentive to imitate the reliable agent. It is crucial that the unreliable agent has no incentive to acquire and abuse the principal's trust in the second trust game via imitating the reliable agent's hostage posting decision and honoring trust in the first trust game (again, the equilibrium strategy of the principal would of course not be optimal should imitation pay off for the unreliable agent). Condition (9)

[13] The 'out-of-equilibrium' behavior is easily guessed for both equilibria. In the pooling equilibrium, the principal does not place trust in the first trust game if no hostage has been posted. The unreliable agent would abuse trust also in the first trust game should trust be placed without prior hostage posting. In the second trust game, the principal never places trust, irrespective of earlier moves of the players, while the (unreliable) agent always abuses trust. In the separating equilibrium, the unreliable agent would abuse trust in both trust games and the principal withholds trust in the second trust game if she receives payoff S_1 after the first trust game.

is sufficient in these respects. According to (9), it is more attractive for the unreliable agent that the principal withholds trust in both trust games than to honor trust after hostage posting in the first trust game and subsequently to abuse trust in the second trust game (which is of course still more attractive than posting a hostage, abuse trust already in the first trust game, and receiving no trust from the principal in the second trust game). Note that it is again not necessary for a separating equilibrium that the hostage is binding. However, note also that condition (9) for a signalling effect of hostage posting is more restrictive than conditions (7) and (8): the trust game is played twice and hence incentives for imitating the reliable agent increase for the unreliable agent. Thus, a separating equilibrium in HTGI according to Theorem 2 is not sufficient to produce a 'transfer-effect' of hostage posting for a new trust game.

5. Conclusions

In this paper, we have argued that Axelrod's analysis of conditions for cooperation should be supplemented with analyses of other social conditions and mechanisms that mitigate incentives for non-cooperative behavior ('greed') or that reduce the 'fear' of non-cooperative behavior by partners. We have shown that hostage posting can contribute to the solution of trust problems. A new feature of our analysis has been that this can be due not only to the 'binding' effect of a hostage but may also result from 'signalling' properties of a hostage. A hostage not only reduces the incentive for opportunistic behavior. Rather, in a context with incomplete information and bounded rationality, posting a hostage may signal that opportunistic behavior is not feasible or that an agent has no incentives at all to behave opportunistically. Hence, hostages can contribute to the 'definition of the situation' and to 'framing' among rational actors.

Obviously, the new implications of the model outlined here can be *tested experimentally* in a stringent and simple way. An advantage of an experimental test would be the feasibility of manipulating the probability of meeting an unreliable partner: a random device could determine the 'type' of the agent (his behavioral alternatives or his incentives) at the beginning of the experiment in such a way that the principal knows the relevant probabilities but cannot observe the outcome produced by the random mechanism.

Using the model, we can likewise *derive predictions for social situations outside the laboratory.* Consider once more the specialist and the law firm. First, predictions might refer to characteristics of associates or partners who are ready to commit themselves by moving. Ceteris paribus, we expect that associates or partners tend to move if their transaction costs of moving are low. This will be more likely for associates or partners who had a rented flat compared to those with privately owned real estate. Likewise, associates or partners who are single or who have a household with a partner who is not active on the labor market will be more likely to move than those who have a household with a partner who is active on the labor market. Such predictions might seem trivial. Less trivial are predictions based on the effects of moving on subsequent behavior of the lawyer and the law firm. We expect ceteris paribus that the actual period of employment of those who move tends to be longer, that the law firm will invest more resources in them and, hence, that they will produce more output and will be more influential within the law firm.

Finally, consider *policy recommendations for employers* that follow from our analysis. We are interested in policy recommendations with two properties. First, they should improve the conditions for a pooling or a separating equilibrium. Thus, they should facilitate hostage posting of employees by (a) reducing their transaction costs and (b) increasing the value of the hostage 'moving' for an employee. Second, and simultaneously, these recommendations should be attractive for employers in the sense of (c) economizing on their costs and (d) not requiring agreements with other potential employers that might be difficult to enforce due to coordination problems or due to competition between employers for scarce resources on the labor market. We have already mentioned that employers often reimburse employees for moving costs, thus reducing employees' transaction costs associated with moving. Moreover, employers often reimburse employees for a certain period of time for their travel expenses between their new workplace and their old residence. The idea underlying the latter arrangement seems to be to facilitate the transition period before moving. A typical arrangement seems to be that the reimbursement for moving costs has to be repaid if the employee quits prematurely (e.g., within a period of two or three years) while in such a case employees do not have to repay their reimbursements for travel expenses. The recommendation is rather obvious: increase the value of the hostage 'moving' and reduce transaction costs associated with moving by changing the 'mix' of reimbursements for moving and travel expenses. See to it that reimbursements for travel expenses likewise have to be repaid if the employee quits prematurely or take care that these reimbursements are paid only after the employee has actually moved. Also, increase reimbursements for moving while decreasing the size of reimbursements for travel expenses or the length of the period for which reimbursements for travel expenses are available.

We would like to close with a more 'philosophical' remark. Via hostage posting a player manipulates his own outcomes in situations with strategic interdependence. Our analysis shows that it can be individually rational to post hostages. Imagine now that a player is not only able to manipulate his outcomes but also to directly manipulate his preferences over outcomes. Note that it follows directly from the analysis presented here that rational actors being able to choose and modify their own preferences would be willing to do so in social dilemmas like trust games (see Hegselmann/Raub/Voss 1986 and Raub/Voss 1990 for a related analysis of endogenous preference changes).

Bibliography

Aumann, R. J. (1981), Survey of Repeated Games, in: R. J. Aumann et al., *Essays in Game Theory and Mathematical Economics in Honor of Oskar Morgenstern*, Mannheim, 11–42

Axelrod, R. (1984), *The Evolution of Cooperation*, New York

— /W. D. Hamilton (1981), The Evolution of Cooperation, in: *Science 211*, 1390–1396

Becker, G. S. (1991), *A Treatise on the Family*, enlarged ed., Cambridge

Binmore, K. (1998), *Game Theory and the Social Contract II. Just Playing.* Cambridge

Blau, P. M. (1964), *Exchange and Power in Social Life*, New York

Buskens, V. (1999), *Social Networks and Trust*, Amsterdam

— /J. Weesie (1999), Cooperation via Social Networks, this volume

Calvert, R. L. (1992), Rational Actors, Equilibrium, and Social Institutions, *mimeo*, Rochester

Camerer, C./K. Weigelt (1988), Experimental Tests of a Sequential Equilibrium Reputation Model, in: *Econometrica 56*, 1–36

Coleman, J. S. (1990), *Foundations of Social Theory*, Cambridge

Colman, A. (1982), *Game Theory and Experimental Games*, Oxford

Craswell, R. (1993), On the Use of 'Trust': Comment on Williamson, 'Calculativeness, Trust, and Economic Organization', in: *Journal of Law and Economics 36*, 487-502

Dasgupta, P. (1988), Trust as a Commodity, in: D. Gambetta (ed.), *Trust: Making and Breaking Cooperative Relations*, Oxford, 49–72

Esser, H. (1993), *Soziologie. Allgemeine Grundlagen*, Frankfurt

Granovetter, M. (1985), Economic Action and Social Structure: The Problem of Embeddedness, in: *American Journal of Sociology 91*, 481–510

Hammerstein, P./R. Selten (1984), Gaps in Harley's Argument on Evolutionarily Stable Learning Rules and in the Logic of 'Tit for Tat', in: *Behavioral and Brain Sciences 7*, 115–116

Harsanyi, J. (1967–8), Games with Incomplete Information Played by 'Bayesian' Players, in: *Management Science 14*, 159–182, 320–324, 486–502

Hegselmann, R./W. Raub/Th. Voss (1986), Zur Entstehung der Moral aus natürlichen Neigungen. Eine spieltheoretische Spekulation, in: *Analyse & Kritik 8*, 150–177

Hirshleifer, D./E. Rasmusen (1989), Cooperation in a Repeated Prisoner's Dilemma with Ostracism, in: *Journal of Economic Behavior and Organization 12*, 87–106

Kelley, H. H./J. W. Thibaut (1978), *Interpersonal Relations. A Theory of Interdependence*, New York

Kreps, D. M. (1990), Corporate Culture and Economic Theory, in: J. E. Alt/K. A. Shepsle (eds.), *Perspectives on Positive Political Economy*, Cambridge, 90–143

— /P. Milgrom/J. Roberts/R. Wilson (1982), Rational Cooperation in the Finitely Repeated Prisoner's Dilemma, in: *Journal of Economic Theory 27*, 245–252

Lindenberg, S. (1992), An Extended Theory of Institutions and Contractual Discipline, in: *Journal of Institutional and Theoretical Economics 148*, 125–154

— (1994), Norms and the Power of Loss: Ellickson's Theory and Beyond, in: *Journal of Institutional and Theoretical Economics 150*, 101–113

Mlicki, P. P. (1996), Hostage Posting as a Mechanism for Cooperation in the Prisoner's Dilemma Game, in: W. B. G. Liebrand/D. M. Messick (eds.), *Frontiers in Social Dilemma Research*, Berlin, 165–183

— /C. Snijders (1995), Two Types of Commitment and the Prevention of Conflict, ISCORE paper 41, *mimeo,* Utrecht

Nash, J. (1951), Non-cooperative Games, in: *Annals of Mathematics 54*, 286–295

North, D. C. (1990), *Institutions, Institutional Change and Economic Performance*, Cambridge

Rapoport, A. (1974), Introduction, in: A. Rapoport (ed.), *Game Theory as a Theory of Conflict Resolution*, Dordrecht, 1–14

— /A. M. Chammah (1965), *Prisoner's Dilemma*, Ann Arbor

Rasmusen, E. (1994), *Games and Information: An Introduction to Game Theory*, 2nd edition, Oxford

Raub, W./G. Keren (1993), Hostages as a Commitment Device: A Game-theoretic Model and an Empirical Test of Some Scenarios, in: *Journal of Economic Behavior and Organization 21*, 43–67

— /Th. Voss (1990), Individual Interests and Moral Institutions. An Endogenous Approach to the Modification of Preferences, in: M. Hechter/K.-D. Opp/R. Wippler (eds.), *Social Institutions: Their Emergence, Maintenance and Effects,* New York, 81–117

— /J. Weesie (1990), Reputation and Efficiency in Social Interactions: An Example of Network Effects, in: *American Journal of Sociology 96*, 626–654

— / — (1992), The Management of Matches. Decentralized Mechanisms for Cooperative Relations with Applications to Organizations and Households, ISCORE paper 1, *mimeo,* Utrecht

— / — (1993), Symbiotic Arrangements: A Sociological Perspective, in: *Journal of Institutional and Theoretical Economics 149*, 716–724

Schelling, T. C. (1960), *The Strategy of Conflict*, London

Schotter, A. (1981), *The Economic Theory of Social Institutions*, Cambridge

Schuessler, R. (1989), Exit Threats and Cooperation under Anonymity, in: *Journal of Conflict Resolution 33*, 728–749

Selten, R. (1965), Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit, in: *Zeitschrift für die gesamte Staatswissenschaft 121*, 301–324, 667–689

— (1975), Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games, in: *International Journal of Game Theory 4*, 25–55

Snijders, C. (1996), *Trust and Commitments*, Amsterdam

Spence, A. M. (1974), *Market Signaling: Information Transfer in Hiring and Related Processes*, Cambridge

Vanberg, V./R. Congleton (1992), Rationality, Morality, and Exit, in: *American Political Science Review 86*, 418–431

Weesie, J. (1992), Disciplining via Exit and Voice, ISCORE paper 88, *mimeo,* Utrecht

— (1994), Social Orientations in Symmetric 2 × 2 Games. Theoretical Predictions and Empirical Evidence, ISCORE paper 17, *mimeo,* Utrecht

Weesie, J./W. Raub (1996), Private Ordering: A Comparative Institutional Analysis of Hostage Games, in: *Journal of Mathematical Sociology 21*, 201–240

Williamson, O. E. (1985), *The Economic Institutions of Capitalism*, New York