university of groningen

University Medical Center Groningen

University of Groningen

# Commitments by hostage posting

Raub, W.

*Published in:*
Perspectives in Moral Science

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*
Publisher's PDF, also known as Version of record

*Publication date:*
2009

Link to publication in University of Groningen/UMCG research database

*Citation for published version (APA):*
Raub, W. (2009). Commitments by hostage posting. In M. Baurmann, & B. Lahno (Eds.), *Perspectives in Moral Science* (pp. 207-225).

**Copyright**
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

**Take-down policy**
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.*

Download date: 13-02-2023

*RMM Vol. 0, Perspectives in Moral Science,*
*ed. by M. Baurmann & B. Lahno, 2009, 207–225*
*http://www.rmm-journal.de/*

*Werner Raub*

# Commitments by Hostage Posting*

**Abstract:**
We survey research on incurring commitments by voluntary hostage posting as a mechanism of cooperation. The Trust Game is employed as a paradigmatic example of cooperation problems. We sketch a very simple game-theoretic model that shows how voluntary hostage posting can bind the trustee and thus induce trustfulness of the trustor as well as trustworthiness of the trustee. We then indicate how the model can be improved by including uncertainty and incomplete information, transaction costs of hostage posting and compensating effects as well as signaling effects of hostages. Further extensions of the theoretical analysis are outlined as well as testable hypotheses and references to empirical research. Problems for future research are suggested.

## 1. Why Consider Commitments by Hostage Posting as a Mechanism of Cooperation?

The study of cooperation by 'rational egoists' goes back, at least, to the $17^{th}$ and $18^{th}$ century philosophers and social scientists Hobbes and Hume. In fact, Parsons (1937, 91) argued that the problem of social order constitutes "the most fundamental empirical difficulty of utilitarian thought". Quite a bit of Hartmut Kliemt's own research focuses on this problem, including early and original contributions such as his discussion of Taylor's (1976) seminal analysis (Kliemt and Schauenberg 1982) and of course his two impressive monographs (Kliemt 1985; 1986) that provide a comprehensive account both through an overview and systematic reconstruction of the literature as well as through theoretical modeling of conditions for cooperation using game-theoretic tools. Arguably, at least in Germany, Kliemt and Voss (1982; 1985) have been the first in social philosophy and in sociology, who clearly recognized and applied the potential of game-theoretic tools for tackling Parsons' challenge.

Kliemt's early work was on cooperation through repeated encounters between the same actors (see also Axelrod 1984 who offered an approach that was

* Support by the Netherlands Organization for Scientific Research (NWO) for the PIONIER-program *The Management of Matches* (grants S 96-168 and PGS 50–370) and for the project *Commitments and Reciprocity* (grant 400–05–089) is gratefully acknowledged. The paper draws on own earlier work on commitments and hostage posting (see references) and I would like to acknowledge the contributions of my previous coauthors Jeroen Weesie, Thomas Voss, Chris Snijders, Gideon Keren, and Vincent Buskens.

in various respects similar to Taylor's from almost a decade earlier but was much more successful in coming up with an easily accessible account of the underlying mechanisms). An extension are encounters in a network of actors with information exchange about others' behavior (Raub and Weesie 1990). Dyadic embeddedness in the sense of repeated encounters and network embeddedness in the sense of encounters and information exchange with third parties allow for cooperation through learning and control (Buskens and Raub 2002; 2010). Learning is based on past interactions. Actors can learn about their partners from own previous encounters with those partners or from information about encounters of third parties with those partners, thus providing opportunities and incentives for cooperation based on reputation building (e.g., Kreps et al. 1982; Buskens 2003). Control is based on the prospect of future interactions. Given the prospect that actors meet again, conditional cooperation can become feasible for rational egoists because others' cooperation today can be rewarded by own cooperation in the future, while others' opportunism today can be punished by withholding own cooperation in the future.

Dyadic embeddedness and network embeddedness are sometimes, i.e., under the right set of conditions such as a sufficiently high likelihood of future encounters, but not always sufficient for cooperation among rational egoists. Extreme cases are one-shot encounters and no exchange of information between actors as well as situations with a 'golden opportunity' for opportunistic behavior in the sense that opportunism is particularly attractive so that the prospect of future sanctions is an insufficient incentive. This raises the problem of specifying alternative or complementary mechanisms of cooperation.

Commitments by voluntary hostage posting in the sense of pledging a bond can be a mechanism of cooperation even without repeated encounters and networks. The basic idea underlying this mechanism is sharply illustrated with an extreme but instructive example that is due to Schelling. The example concerns a cooperation problem between a kidnapper who got 'cold feet' and her prisoner. The kidnapper must trust her prisoner not to turn to the police immediately after being set free. How to find a safeguard that allows the placement of trust? "Both the kidnapper [...] and the prisoner may search desperately for a way to commit the latter against informing on his captor [...]. If the victim has committed an act whose disclosure could lead to blackmail, he may confess it; if not, he might commit one in the presence of his captor, to create the bond that will ensure his silence." (Schelling 1960, 43f.)

Our own paradigmatic example for cooperation by voluntary hostage posting concerns employment relations and the labor market for professionals. The example seems to have some similarity with typical features of cooperation problems between universities and their senior faculty, at least in the German academic system. Employers often have to decide on making investments in an employee that are largely relationship-specific (see Becker 1964 on investments in human capital). For example, the employer provides training and schooling or she adapts her organization's internal structure and processes to the employee's expertise, supplying him with assistants and additional staff, etc. Much

of these investments have to be depreciated should the employee decide to quit. If the employer invests, the employee may have to decide between maintaining a durable relationship with the employer or using the new appointment as a stepping stone and accepting an outside offer. Assuming that the employer has invested in general human capital of the employee, it seems likely that the employee will have opportunities and incentives for quitting. After all, through his increased general human capital the employee becomes more attractive for other employers who, moreover, do not need to recover the costs of training and schooling. One way of solving this problem is a contract between the employer and the employee stipulating that the employee has to pay at least part of the costs of training and schooling if he quits prematurely. Another way of solving the problem is non-contractual. The employee moves and acquires real estate close to his job. Thus, if he quits prematurely and accepts an offer from another and far away employer, he would have to move again and incur the financial as well as social costs associated with moving.

Hostage posting is a typical example of a 'strategic move' through incurring a commitment in Schelling's (1960) sense. Schelling accentuated the use of hostages and other commitments in bargaining contexts. He showed that a committed actor can frequently induce a bargaining outcome that is favorable for himself and unfavorable for the partner. Thus, counterintuitively, an actor often has incentives to incur a commitment and to bind himself voluntarily: "the power to constrain an adversary may depend on the power to bind oneself [...] in bargaining, weakness is often strength, freedom may be freedom to capitulate, and to burn bridges behind one may suffice to undo an opponent." (Schelling 1960, 22) Other than Schelling, though, we consider how and when commitments such as hostages allow for mutually beneficial adjustments. The hostage modifies subsequent incentives for cooperation. By posting an appropriate hostage in the first place, an actor incurs a commitment that serves as a safeguard for the partner to cooperate (Schelling 1960; Williamson 1985).

Hostage posting becomes feasible under institutional embeddedness of the trust problem (Weesie and Raub 1996). We assume an institutional context as given that provides opportunities for actors to post a hostage ex ante, before they decide on whether or not to cooperate. The context providing the opportunity to post a hostage is considered as an exogenous condition. By using this opportunity and posting a hostage, actors create a private institution. As Coleman (1990) put it, actors create a "constructed social environment" that promotes cooperation. Thus, institutional embeddedness provides opportunities for private ordering (Macaulay 1986; Williamson 1985) of relations. The private institution itself—the hostage—is endogenous. We thus focus on conditions such that these private institutions result from individually rational equilibrium behavior and are therefore self-enforcing (see Schotter 1981 and Calvert 1995 for the distinction between institutions as exogenous constraints and as outcomes of equilibrium behavior). Therefore, we do not assume that an external third party forces actors to incur a commitment by hostage posting but address the deeper question concerning the conditions such that a hostage is posted voluntarily

and without external coercion (it is noteworthy that Kliemt 1986, 332–349) has clearly seen the usefulness of commitments for solving cooperation problems as well as the theoretical challenge to account for how rational egoists can and will incur commitments). Notice that institutional embeddedness includes but is not restricted to the legal infrastructure of social relations and the enforcement of contractual agreements on hostages through the law. Hostages are sometimes posted through a contractual agreement like the contract stipulating that the employee compensates the employer for some of the training costs in case of a premature quit. In other cases, however, hostages are posted informally and the enforcement of the commitments incurred through hostage posting is not provided through the law. The employee who moves close to his job posts a hostage in a non-contractual way.

## 2. A Very Simple Model for Cooperation by Hostage Posting

Consider a meanwhile standard example for cooperation problems, namely, trust problems in the sense that the trustee has incentives to abuse trust and the trustor has something to lose if trust is abused (Coleman 1990, chapter 5). A well-known model of trust problems is the standard Trust Game as introduced by Dasgupta (1988, 59–61) and Kreps (1990, 100–101). The game is played by two actors. Actor 1 is the trustor and actor 2 is the trustee. The trustor moves first—Coleman stresses the importance of this feature: there is a time lag between the action of the trustor and the action of the trustee—and chooses between placing trust 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 trust is not placed, with payoffs (cardinal utility) $P_i$ ($i = 1,2$) for trustor and trustee. If trust is placed, the trustee chooses between honoring and abusing trust. If trust is honored, trustor and trustee receive $R_i > P_i$. If trust is abused, the trustor receives $S_1 < P_1$, while the trustee receives $T_2 > R_2$. The trustee thus has an incentive to abuse trust. Note that these assumptions capture Coleman's argument that placing trust involves a risk. The trustor is better off if trust is placed and honored than if she withholds trust. On the other hand, if the trustee abuses trust, the trustor is worse off compared to the situation when trust is not placed. The trust *problem* is therefore twofold. First, by placing trust, the trustor incurs risks such as trust being abused. Second, if the trustor decides not to place trust, both trustor and trustee could have been better off had trust been placed and honored. In Schelling's example, the kidnapper is in the role of the trustor, while the prisoner is in the role of the trustee. In our own example, the employer is the trustor and the employee is the trustee.[1]

[1] In our examples above as well as throughout the paper, 'she' refers to the trustor, while 'he' refers to the trustee.

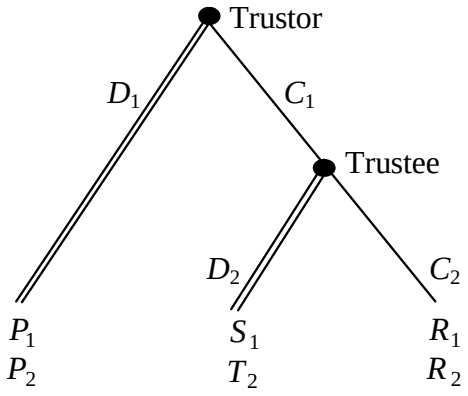


Figure 1: The Trust Game ($S_1 < P_1 < R_1$, $P_2 < R_2 < T_2$)

Here and subsequently throughout this paper, we use standard game-theoretic assumptions (see, e.g., Rasmusen 2007) and assume that both actors know the structure of the game, know that the other actor knows the structure of the game, and so forth. We furthermore assume that the game is played noncooperatively: actors are unable to make enforceable agreements or enforceable one-sided commitments except agreements and commitments explicitly modeled as moves in the structure of the game. Obviously, we assume a noncooperative game precisely because we wish to specify conditions such that rational actors will be prepared to make these agreements and commitments and to specify conditions such that these agreements and commitments will actually induce trustfulness as well as trustworthiness. Thus, we employ the Nash program (Nash 1951) of explicitly modeling bargaining, communication, and all other kinds of pre-play behavior as moves in an extended noncooperative game, and to derive cooperative behavior as an equilibrium of that extended game.

The Trust Game has a unique subgame perfect equilibrium such that the trustor withholds trust while the trustee would abuse trust (indicated by the double lines in Figure 1). This equilibrium is Pareto-suboptimal since both actors are better of when trust is placed and honored and, hence, the Trust Game models cooperation problems. In fact, the Trust Game can be seen as a one-sided version of the Prisoner's Dilemma: one and only one actor, the trustee, has opportunities and incentives for opportunistic behavior.

To specify conditions for posting a hostage and for generating and stabilizing trust by hostage posting, we introduce a Hostage Game which is an extended version of the Trust Game. In the Hostage Game, the trustee moves first and chooses between posting or not posting a hostage. Subsequently, the trustor is informed on the trustee's initial move and the actors play the Trust Game. Note that the trustor can thus condition the placement of trust on the hostage posting

decision of the trustee. Assume that the trustee loses his hostage if and only if he posts it at the beginning of the game, if the trustor places trust, and if the trustee subsequently abuses trust. The hostage has value $K_2 > 0$ for the trustee and we assume that the trustee's utility at the end of the Hostage Game is additive in his payoff at the end of the corresponding Trust Game and the value of a possibly lost hostage. For simplicity, we also assume for the time being that hostage posting is not associated with transaction costs for the trustee and that a hostage that is lost by the trustee is not given to the trustor (or, equivalently, the hostage has no value for the trustor).

One easily verifies (Weesie and Raub 1996, 218) that the Hostage Game has a subgame perfect equilibrium such that the trustee posts a hostage, the trustor places trust, and the trustee honors trust if $K_2 > T_2 - R_2$. This confirms the intuition that it can be individually rational for the trustee to voluntarily post a hostage and that the hostage can be a sufficient safeguard for the trustor to place trust. This is the case if the hostage is sufficiently valuable for the trustee so that he has incentives to honor trust if trust is placed after hostage posting. The critical value that makes the hostage binding are the trustee's costs $T_2 - R_2$ of honoring trust in the Trust Game.

Note that the equilibrium strategies that induce hostage posting, placement of trust, and honoring trust are in one respect similar to conditional cooperation in repeated encounters. Namely, the equilibrium strategies are 'reactive' in the sense that both actors condition own behavior on the behavior of the partner. The equilibrium strategy of the trustor makes placement of trust dependent on prior hostage posting of the trustee. Her equilibrium strategy comprises the tacit promise to place trust after hostage posting and the tacit threat to withhold trust if the trustee does not post trust. Conversely, the trustee's equilibrium strategy comprises the tacit promise not to abuse the trustor's trust. Subgame perfectness of the equilibrium makes for credibility of these threats and promises.

## 3. Improving the Model

The simple Hostage Game, while instructive, also suffers from various weaknesses of which we mention five. First, transaction costs of hostage posting have been assumed away. In our labor market example, though, acquiring real estate close to the new job and moving comes with considerable transaction costs for the employee: 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 employee, his partner, and children that are tied to his former place of residence. Also, the employee incurs costs by undermining his bargaining position should a new outside offer emerge.

Second, we have assumed that the trustee's hostage has no value for the trustor. However, hostage posting by the employee through a contract stipulating that the employee has to pay at least part of the costs of training and

schooling if he quits prematurely clearly involves not only a binding effect for the employee that makes accepting a new job less attractive. Such a contract likewise involves a compensating effect for the trustor that reduces her damage in case of opportunistic behavior of the trustee (Williamson 1985, 172 likewise highlighted the distinction between the value of the hostage for the actor who posts the hostage and the value of the hostage for the other party).

Third, our model neglects incomplete information problems. Not all employees have or can expect an outside offer. Typically, there will be asymmetric information between the employer and the employee in the sense that the employee knows whether or not he has or can expect such an offer while the employer at best can derive the probability for opportunities and incentives for a premature quit from observable characteristics of the labor market as well as from observable characteristics of the employee himself. Under such circumstances, the hostage may serve as a signal for the trustor about unobservable characteristics of the trustee that are related to his opportunities and incentives for abusing trust. The move of the employee close to his job may then signal that he does not have and does not expect an outside offer (our distinction between reducing the incentives of the trustee for abusing trust through the binding effect of hostage posting and using hostages as signals for unobservable characteristics of the trustee is similar to the discussion of 'ex post' bonding and 'ex ante' screening effects of hostages in Williamson 1985, 168).

Fourth, our model also neglects problems due to uncertainty. For example, the employer's investments in her relation with the new employee may be lost not because the employee quits opportunistically but because the employee falls ill.

Finally, our model includes not only problematic assumptions but also has at least one implication that seems highly problematic. Namely, the model implies that the trustee would be willing to post a hostage even if such a hostage is extremely valuable for him. This, however, seems intuitively less than plausible and empirically the value of hostages seems to be typically limited. For example, while contracts are not uncommon that stipulate that an employee has to provide compensation for training costs in the case of a premature quit, the compensation covers only some of those costs and is in any case not excessively high.

Fortunately, remedying these weaknesses simultaneously is feasible. Raub (2004) provides an extended model of trust by hostage posting with the following features:

- Uncertainty is included in the underlying Trust Game in the sense that 'things may go wrong' (e.g., the employee falls ill) due to unfavorable external contingencies. The possibility that things can go wrong is accounted for by assuming a chance event (a 'move of Nature') after trust has been placed and honored.
- Incomplete information of the trustor about the trustee is modeled by assuming that the trustee has opportunities and incentives for abusing trust not with certainty but only with some positive probability. The trustor only

knows the probability such that the trustee has opportunities and incentives to abuse trust but does not know the actual behavioral alternatives and incentives of the trustee. The trustor can meet two possible 'types' of trustees. The first type is 'unreliable' in the sense that he has opportunities and incentives to abuse trust. If the trustor places trust, such a trustee can either honor ($C_2$) or abuse trust ($D_2$). A second type of trustee is 'reliable' in the sense that he has no opportunities for abusing trust. If the trustor meets such a trustee, the trustee 'automatically' honors trust and the game ends after trust has been placed. A chance move of Nature at the beginning of the game determines the type of trustee playing the game. The outcome of Nature's initial move and hence the trustee's actual type is known to the trustee himself but is unobservable for the trustor. Such a Trust Game with incomplete information provides room for possible signaling effects of hostage posting.[2]

- In the extended model, the Trust Game with incomplete information is again embedded in a Hostage Game: Nature moves first and determines the trustee's type. Nature's move is observed by the trustee but the trustor cannot observe the trustee's type and is only informed on the probabilities for either type of trustee. Subsequently, the trustee moves and chooses between posting or not posting a hostage. The trustor is informed on the trustee's hostage posting decision. Afterwards, the trustor decides to place ($C_1$) or to withhold trust ($D_1$). The game ends after $D_1$. If the trustee happens to be of the reliable type who cannot abuse trust, the game also ends after $C_1$ (more precisely, the game ends after the reliable trustee's 'trivial' move $C_2$ and the subsequent chance move of Nature that determines whether or not 'things go wrong' although the trustee honors trust). If the trustor places trust and the trustee is unreliable, the trustee chooses between honoring ($C_2$) and abusing trust ($D_2$) and the game ends (again, if trust is honored, after the chance move of Nature that determines if 'things go wrong').
- In the extended model it is assumed that the trustee loses his hostage if and only if he posts it, the trustor places trust, and the trustee either abuses trust or honors trust but unfavorable contingencies obtain and 'things go wrong' even though the trustee abstains from opportunistic behavior. Hostage posting is thus considered as a mechanism to mitigate risks from opportunistic

[2] Alternative assumptions are conceivable. For example, one could imagine that a reliable trustee has opportunities to abuse trust but has no incentives to do so. This would be the case if the trustee has internalized norms and values inducing sufficient 'internal sanctions' should he abuse trust and behave opportunistically. Then, honoring trust would be associated with a higher 'net utility' than abusing trust (see Camerer and Weigelt 1988; Dasgupta 1988; Bacharach and Gambetta 2001; and Güth and Ockenfels 2003 for such models of trust games with incomplete information). Such an alternative scenario leads to similar results like the one considered here. We prefer our conceptualization because it seems to fit better with a rational choice approach by focusing on the possible lack of opportunities to abuse trust rather than the possibility that the trustee is '(over)socialized': we keep our assumptions on individuals and their preferences as simple as possible while complexity is introduced into the model via assumptions on their interaction situation such as assumptions on feasible actions and restrictions (see, e.g., Wippler and Lindenberg 1987; Coleman 1990, passim).

behavior of the trustee as well as risks from uncertainty. For simplicity, we assume that the value of the hostage for the reliable trustee is the same as the value for the unreliable trustee.

- Hostage posting is associated with transaction costs for the trustee in the extended model. We consider transaction costs in Williamson's (1985, 20–22, 388) sense, that is, costs of drafting, negotiating, setting up, and running the hostage arrangement. Examples are opportunity costs due to the temporary impossibility of using the hostage, enforcement costs through hiring a lawyer who monitors the hostage arrangement, the risk that a hostage posted under the control of the partner is unexpectedly not returned later on, and costs that results from reduced flexibility to respond to changes in exogenous conditions. These costs arise if and only if the trustee decides to post the hostage and do not depend on how the game develops after hostage posting. Hence, these costs are not only due if the trustee loses his hostage (in this case, one could consider the transaction costs simply as an ingredient of the value $K_2$ of the hostage for the trustee) but also if trust is placed and honored after hostage posting as well as if no trust is placed after hostage posting.
- A core assumption of the extended model is that the transaction costs of hostage posting can differ between the two types of trustees. On the one hand, allowing for differences in transaction costs seems realistic. In the example of the employee who posts a hostage by moving close to his job, an important ingredient of the transaction costs involved in posting the hostage is due to undermining the employee's bargaining position vis-à-vis an alternative employer offering a new position. These transaction costs emerge by definition for an employee who is unreliable in the sense of our model, while a reliable trustee does not have to incur these costs. While the assumption that transaction costs of hostage posting can differ between different types of trustees seems to be empirically plausible, differences in transaction costs are also interesting from our theoretical perspective on trust based on hostage posting. After all, one expects from signaling theory (Spence 1974) that the signaling function of hostages depends on differences in signaling costs for different types of trustees. Transaction costs associated with hostage posting can be conceived as signaling costs of hostage posting and therefore the question arises if differences in transaction costs for different types of trustees can affect the signaling function of hostages.
- To account for the compensation function of hostages, the extended model assumes that a lost hostage is given to the trustor for whom the hostage has value $K_1$. The case $K_1 = 0$ can be interpreted as the situation such that a lost hostage is not given to the trustor (see Weesie and Raub 1996, 214–216, for a detailed discussion of 'hostage institutions', that is, exogenous rules of the hostage game that determine under what conditions actors lose a hostage and what happens with a lost hostage). Note that our examples differ with respect to the value of the hostage for the trustor. The contract stipulating that the employee has to pay for (some) training costs in case of a prema-

ture quit is a relatively valuable hostage for the employer. The value of the hostage 'moving close to one's job', while being high for the employee who posts the hostage, is low for the employer.

- In the extended model, each actor's payoff can be assumed to be additive in the actor's payoff at the end of the corresponding Trust Game, in the value of a hostage lost or received, and in transaction costs of hostage posting (see Raub 2004, 339, for the extensive form of the game).

The extended model allows for deriving generic conditions such that placing and honoring trust based on hostage posting becomes individually rational in the sense of being supported by a perfect Bayesian equilibrium (see Rasmusen 2007 for a discussion of the concept). For appropriate values of the model parameters, a pooling equilibrium exists, i.e., an equilibrium such that the hostage posting decision of both types of trustees is the same so that hostage posting does not signal the type of the trustee (Raub 2004, Theorem 1). For other parameter values, the model implies the existence of separating equilibria of the extended model (Raub 2004, Theorem 2). These are equilibria such that the two types of trustees differ with respect to their hostage posting decision so that hostage posting involves a signal for the trustor about the type of the trustee. The interesting case from a substantive perspective involves that the reliable trustee posts a hostage while the unreliable trustee chooses not to post a hostage and the trustor places trusts after hostage posting but withholds trust if no hostage has been posted. Here, posting a hostage signals that the trustee is reliable. In our example, this would be the case such that the employee's willingness to move close to his job would indicate that he does not have and does not expect to receive an outside offer from another employer. Rather than providing technical details, we summarize the substantive conditions for trust based on hostage posting that follow from the analysis of the extended model.

The extended model implies, *first*, that the hostage has to be valuable enough for the trustee if the hostage promotes trust by binding the trustee. Binding is an issue if the trustee who posts the hostage may be unreliable—the reliable trustee cannot abuse trust. The theorem on the pooling equilibrium (a hostage is posted by the unreliable trustee) specifies a lower bound on the value $K_2$ of the hostage for the unreliable trustee and this lower bound depends on the unreliable trustee's costs $T_2 - R_2$ of honoring trust in the Trust Game.

*Second*, the hostage has to be valuable enough for the trustor for inducing trust through compensation. The extended model allows to specify a lower bound on the value $K_1$ of the hostage for the trustor. The lower bound on the value of compensation depends crucially on the amount of uncertainty that is associated with placing trust. If the risk of unfavorable contingencies is high, the hostage must provide positive compensation for the trustor. If the probability of unfavorable contingencies is small enough, it suffices that the hostage binds the trustee or signals that the trustee is reliable and it is not required that the hostage also includes positive compensation of the trustor in case things go wrong.

The *third* result is that signaling via hostages is related to differences between different types of trustee in their transaction costs associated with hostage posting. In the extended model, differences in transaction costs are the only possible differences between the two types of trustees. The extended model implies that signaling requires that the transaction costs of hostage posting are lower for the type of trustee who posts the hostage than for the other type of trustee who does not post the hostage.

The *fourth* condition for trust based on hostage posting that follows from the extended model is that transaction costs associated with hostage posting have to be small enough. An equilibrium such that a hostage is posted requires upper bounds on the transaction costs of hostage posting for the (type of) trustee who posts the hostage. More precisely, the upper bound depends on the gains $R_2 - P_2$ from honored trust compared to the situation of withheld trust. Also, the upper bound on the transaction costs of hostage posting depends on the probability of unfavorable contingencies after trust has been placed and honored.

*Fifth*, the extended model implies that the hostage should not be too valuable for the trustee in order to allow for trust via hostage posting. The extended model provides upper bounds on the value $K_2$ of the hostage for a trustee who posts the hostage. The upper bound depends on the gains $R_2 - P_2$ from honored trust compared to the situation of withheld trust. One can show that an upper bound on the value of the hostage for the trustee is only required if the underlying trust problem involves uncertainty so that the trustee may lose his hostage even without opportunistic behavior.

A *sixth* result concerns properties of the equilibrium strategies that induce hostage posting as well as placing and honoring trust. Just like in the simple model from section 2, the equilibrium strategies are 'reactive'. The strategy of the trustor makes her placement of trust dependent on prior hostage posting of the trustee. Her strategy carries the tacit promise to place trust after hostage posting and the tacit threat to withhold trust if the trustee does not post a hostage. Conversely, if the unreliable trustee posts a hostage in equilibrium, his strategy implies a tacit promise not to abuse the trustor's trust. The equilibrium ensures that these threats and promises are credible.

*Finally*, one can show that a pooling or a separating equilibrium inducing trust by hostage posting is either unique or is at least a weak Pareto-improvement compared to all other—if any—equilibria in the extended model. Thus, if such equilibria exist, equilibrium selection problems are not severe. The prediction that rational actors will post hostages and will place and honor trust follows from the uniqueness of the equilibrium or from payoff dominance arguments.

It is helpful to relate these results to our labor market example example of a trust problem and of hostages as a mechanism to induce trust. A contract stipulating that the employee has to reimburse the employer if he quits prematurely presumably serves binding purposes and also provides some compensation for the employer. The hostage 'moving and acquiring real estate close to the job' does not provide compensation for the trustor. This hostage seems to serve not only binding but also signaling purposes. After all, the transaction costs of post-

ing the hostage differ severely between employees who have or expect an outside offer and those who do not have and do not expect such an offer.

As usual, the question arises on whether the results generalize to other cooperation problems than 'only' trust problems like in our models. E.g., do results generalize to other 2- and $n$-person social dilemma games? Do results generalize to other 'hostage institutions', i.e., rules that specify under what conditions a hostage is lost, what happens to a lost hostage, and also specify the conditions under which transaction costs of hostage posting arise and for whom those costs arise. Weesie and Raub (1996) provide quite some such generalizations, albeit in complete information contexts.

## 4. Testable Implications and Policy Recommendations

The extended model implies testable predictions for *laboratory experiments*. Obviously, such experiments require explicit assumptions on how material incentives provided by the experiment relate to utility functions of subjects. In an experimental design, the extended model could be implemented by using a random device that determines the type of the trustee at the beginning of the experiment in such a way that the trustor knows the relevant probabilities but cannot observe the outcome produced by the random mechanism. An approach for deriving predictions for experiments is to vary some parameter of the extended model that is interesting from a substantive perspective so that different model implications on trust via hostage posting apply. To illustrate, consider varying the probability of unfavorable contingencies. For example, compensation of the trustor through hostage posting in the sense that the hostage is rather valuable for the trustor should have a stronger effect on hostage posting as well as on inducing trust by hostage posting if the underlying trust problem involves a high probability of unfavorable contingencies. Research in this direction seems to be particularly promising because much experimental evidence on hostages as a mechanism of cooperation in dilemmas such as the Trust Game and the Prisoner's Dilemma (Raub and Keren 1993; Snijders 1996, chapter 6; and Snijders and Buskens 2001) indicates that—empirically—positive compensation through hostage posting has a strong effect on trust and cooperation.

A focus on the effects of uncertainty and the probability of unfavorable contingencies seems also useful for *empirical applications outside the laboratory*. The prediction would be that hostages tend to provide more compensation for the trustor if the underlying trust problem is one with a higher probability of unforeseen contingencies. Conversely, keeping the probability of unforeseen contingencies constant, one would expect that binding and signaling properties of hostages are stronger ($K_2$ increases and different types of trustees differ strongly with respect to the transaction costs of hostage posting) in case of increasing risks from opportunistic behavior. Other testable predictions can be generated by considering the effects of increasing differences in transaction costs of hostage posting for different types of trustees. Based on our theorem on the separating equilibrium,

we would expect that increasing differences in transaction costs come with more variation in hostage posting decisions. Finally, predictions for social situations outside the laboratory could focus on characteristics of trustees who are willing to post certain types of hostages. For example, an employee will tend to move and acquire real estate close to his new job if his transaction costs associated with posting this hostage are low: the employee will be more likely to move if he had a rented flat rather than privately owned real estate, if he is single or has a household with a partner who is not active on the labor market. While predictions of such characteristics may seem obvious, predictions on the effects of hostage posting for subsequent behavior of trustor and trustee may be more interesting. One would expect that the actual period of employment is longer if the employee moves, that the investments of the employer in the employee will be higher, and that the employee will produce more output and will be more influential within the employer's firm.

What about *policy recommendations* such as recommendations for employers on how to design the contractual relation with a newly hired professional? Such policy recommendations should have two properties. First, they should improve the conditions for a pooling or a separating equilibrium inducing trust by hostage posting. Thus, they should facilitate hostage posting of employees by reducing their transaction costs or increasing the value of the hostage 'moving'. Second, and simultaneously, these recommendations should be attractive for employers in the sense of economizing on their costs and 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. 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 employees quits prematurely (e.g., within a period of two or three years), while in such a case employees do not have to repay their reimbursement for travel expenses. The recommendation would then be to 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.

## 5. Commitments by Hostage Posting in a Broader Research Program on Mechanisms of Cooperation

Research on commitments by hostage posting can and should be seen as embedded in a broader research program on mechanisms of cooperation. We briefly sketch three perspectives on commitments by hostage posting that emerge from such a broader program.

The first perspective is of a more 'philosophical' nature. Via hostage posting, an actor manipulates his own outcomes in situations with strategic interdependence such as trust problems and other cooperation problems. The analysis shows that it can be individually rational for selfish actors to post hostages. Imagine now that an actor is not only able to manipulate his outcomes but also to directly manipulate preferences over outcomes, with outcomes as such remaining the same. As a variation on the same theme, imagine a third party whose preferences coincide with those of the trustee but who is also able to modify the trustee's preferences—a scenario that is somewhat similar to parents as third party who try to modify their children's preferences through various socialization efforts. It follows directly from the analyses sketched above that rational and selfish actors being able to choose and modify their own preferences would be willing to do so in social dilemmas like Trust Games (see Raub and Voss 1990 for a related analysis of endogenous preference changes, while Güth and Kliemt 2000 approach this problem from an evolutionary angle).

Second, note that our analysis centered on the effects of hostage posting for actors' outcomes after a game has been played. In this sense, our analysis focused on hostage posting as a mechanism of cooperation through outcome-based motivations. However, quite some recent rational choice research on cooperation submits that process-based motivations should be considered, too (e.g., Fehr and Schmidt 2006; Vieth 2009). For example, a trustee honors trust not because of the implications of honoring trust for subsequent outcomes but because he considers the trustor's placement of trust as kind behavior that he wishes to reciprocate by being kind himself through honoring trust. From this perspective, posting or not posting a hostage can affect subsequent placement of trust not only through the effects of hostage posting for outcomes of trustor and trustee but also because, e.g., the trustor considers hostage posting as kind behavior of the trustee that she wishes to reciprocate by placing trust, while she considers the trustee's decision not to post a hostage that he could have posted as unkind behavior that she wishes to reciprocate by withholding trust. In a similar vein, the trustee may be induced to honor trust not only because of a motivation to reciprocate kind behavior of the trustor but also because of a desire to be consistent with his own prior hostage posting decision. One easily intuits that such a mechanism may even work with hostages that have small or no consequences in terms of modifying outcomes, i.e., have small or no binding and compensation effects (see Vieth 2009 for interesting theoretical as well as experimental work on hostages from the perspective of process-based motivations).

As a third perspective, note that an analysis of hostage posting as a mechanism of cooperation has been motivated with the observation that repeated encounters and encounters in a network can be insufficient for stabilizing trust. The model, however, is a model *without* these forms of embeddedness rather than a model such that repeated encounters are *too 'weak'* to stabilize trust. See Raub (1992) as well as Weesie et al. (1998) for work accounting for the combined effects of dyadic and network embeddedness and of opportunities for hostage posting through institutional embeddedness for the solution of trust problems.

## 6. Problems for Further Research

Quite some questions on commitments by hostage posting remain for further research. First, consider the "expropriation hazard" (Williamson 1985, 177) and the loss of flexibility (Becker 1991, 12–13) associated with hostage posting. In the models sketched here, only the trustee and not the trustor faces opportunities and incentives for abusing trust. However, this need not be true. Consider the case of hostage posting of the employee by moving close to his job. This hostage is not subject to an expropriation risk since the hostage as such is not valuable for the employer but posting the hostage reduces the flexibility of the trustee if contingencies change. Such changing contingencies can depend on 'moves of Nature' rather than on incentive guided behavior of the employer. An example is the risk that the firm goes bankrupt due to market fluctuations rather than bad management of the employer as owner/manager. However, contingencies may also change due to strategic behavior of the employer after hostage posting of the employee. For example, when hiring the employee, the employer may offer not only investments in future training and schooling but may also promise future general policies that are attractive for the employee. Such promises are often non-contractual and not legally binding or enforceable. If the employee commits himself by moving, he loses flexibility to react to the employer's future deviations from the promises. Expropriation risks and the expected costs of loss of flexibility can be included in the transaction costs associated with hostage posting. This seems satisfactory for risks that depend on moves of Nature rather than strategic behavior of the partner. A more appropriate analysis that takes strategic behavior of the other actor into account would require to model hostage posting as a mechanism of cooperation in more complex dilemma situations with possibilities for opportunistic behavior of more than one actor. Reciprocal hostage posting by both actors is an obvious mechanism for mitigating risks from two-sided opportunistic behavior (see Williamson 1985, chapter 8, for examples; Weesie and Raub 1996 for formal model building using games with certainty and complete information; and Raub and Keren 1993 for experimental results).

The expropriation risk and the loss of flexibility are *problems* associated with hostage posting. One can also imagine possible further *benefits* of hostage posting in addition to promoting trust. One such benefit that merits closer analysis

is that posting the hostage may make the relation between trustor and trustee more productive. For example, the employee who moves close to his job may not only promote investments of the employer in the employee's training and schooling. Moving close to his job may moreover in itself imply an additional contribution to the employee's productivity and job satisfaction since he benefits from reduced commuting time, better (opportunities for) contacts and interaction with colleagues etc. It should be easy to show that 'productivity' of a hostage in this sense implies that the conditions for trust based on hostage posting become less restrictive since posting the hostage also increases the payoffs $R_i$ from trust that is placed and honored. A variant on this theme is that hostage posting of the trustee is associated with costs for the trustor. For example, an employment contract stipulating that the employee has to pay for some training and schooling in case of a premature quit comes with some costs of contracting also for the employer. This would reduce $R_1$ after a hostage has been posted. In such a scenario, the conditions for trust based on hostage posting become more restrictive.

A final feature of our example that has not been modeled explicitly is hostage selection. The employee can post a hostage by moving to his job or by signing a contract stipulating that he pays back training and schooling costs of the employer after a quit. When to choose one or the other of these hostages, when both? It seems reasonable to assume that the transaction costs associated with posting different hostages as well as the value of the hostage will be important factors that affect the trustee's decision but a more thorough analysis seems useful (see Snijders 2000 for first steps in the analysis of the hostage selection problem).

In a contribution honoring Hartmut Kliemt and his work it would be strange at least not to mention that the approach presented here throughout assumed perfect rationality in the sense of game-theoretic equilibrium behavior. Much of Kliemt's earlier work (such as 1982 and 1986 on cooperation problems) likewise employed such assumptions but he has increasingly moved towards bounded rationality approaches (e.g., Güth and Kliemt 2004). A bounded rationality approach to commitments by hostage posting is not only far beyond the scope of the present contribution but is to this author's best knowledge still lacking in the literature (e.g., while Williamson often refers to bounded rationality, also in his work on commitments and hostages, he typically seems to have in mind various kinds of information limitations that are quite different from bounded rationality in the sense of other than equilibrium behavior). One may add, though, that a core ingredient of the idea of cooperation by hostage posting is that actors post hostages with an eye on the effects of hostage postage for *future* behavior of their partners. Thus, less than perfect rationality may suffice to induce hostage posting but it seems that the mechanisms is closely tied to at least some *forward* looking behavior and *anticipation* of the effects of own present behavior on the future behavior of others (or, even though Kliemt does not favor this interpretation, *as if* behavior of this kind).

## References


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

Bacharach, M. and D. Gambetta (2001), "Trust in Signs", in: K. S. Cook (ed.), *Trust in Society*, New York: Russell Sage, 148–184.

Becker, G. S. (1964), *Human Capital*, New York: NBER, Columbia University Press.

— (1991), *A Treatise on the Family*, enlarged ed., Cambridge/MA: Harvard University Press.

Buskens, V. (2003), "Trust in Triads: Effect of Exit, Control, and Learning", *Games and Economic Behavior* 42, 235–52.

— and W. Raub (2002), "Embedded Trust: Control and Learning", *Advances in Group Processes* 19, 167–202.

— and — (2010), "Rational Choice Research on Social Dilemmas: Embeddedness Effects on Trust", forthcoming in R. Wittek, T. A. B. Snijders and V. Nee (eds.), *Handbook of Rational Choice Social Research*, New York: Russell Sage.

Calvert, R. L. (1995), "Rational Actors, Equilibrium, and Social Institutions", in: J. Knight and I. Sened (eds.), *Explaining Social Institutions*, Ann Arbor: University of Michigan Press, 57–94.

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

Coleman, J. S. (1990), *Foundations of Social Theory*, Cambridge/MA: Belknap Press of Harvard University Press.

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

Fehr, E. and K. M. Schmidt (2006), "The Economics of Fairness, Reciprocity and Altruism—Experimental Evidence and New Theories", in: S.-C. Kolm and J. M. Ythier (eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity*, Amsterdam: Elsevier, 615–91.

Güth, W. and H. Kliemt (2000), "Evolutionarily Stable Co-operative Commitments", *Theory und Decision* 49, 197–221.

— and — (2004), "Perfect or Bounded Rationality?", *Analyse & Kritik* 26, 364–381.

— and A. Ockenfels (2003), "The Coevolution of Trust and Institutions in Anonymous and Non-anonymous Communities", *Jahrbuch für Neue Politische Ökonomie* 20, 157–174.

Kliemt, H. (1985), *Moralische Institutionen*, Freiburg: Alber.

— (1986), *Antagonistische Kooperation*, Freiburg: Alber.

— and B. Schauenberg (1982), "Zu M. Taylors Analysen des Gefangenendilemmas", *Analyse & Kritik* 4, 71–96.

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

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

Macaulay, S. (1986) "Private Government", in: L. Lipson and S. Wheeler (eds.), *Law and the Social Sciences*, New York: Russell Sage, 445–518.

Nash, J. (1951) "Non-Cooperative Games", *Annals of Mathematics* 54, 286–295.

Parsons, T. (1937), *The Structure of Social Action*, New York: Free Press.

Rasmusen, E. (2007), *Games and Information: An Introduction to Game Theory*, $4^{th}$ ed., Oxford: Blackwell.

Raub, W. (1992), "Eine Notiz über die Stabilisierung von Vertrauen durch eine Mischung von wiederholten Interaktionen und glaubwürdigen Festlegungen", *Analyse & Kritik* 14, 187–194.

— (2004), "Hostage Posting as a Mechanism of Trust: Binding, Compensation, and Signaling", *Rationality and Society* 16, 319–365.

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

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

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

Schelling, T. C. (1960), *The Strategy of Conflict*, London: Oxford University Press.

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

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

— (2000), "Trust via Hostage Posting", on CD-ROM in J. Weesie and W. Raub (eds.), *The Management of Durable Relations. Theoretical Models and Empirical Studies of Households and Organizations*, Amsterdam: ThelaThesis, 114–116 and 19ff..

— and V. Buskens (2001), "How to Convince Someone That You Can Be Trusted? The Role of 'Hostages'", *Journal of Mathematical Sociology* 25, 355–383.

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

Taylor, M. (1976), *Anarchy and Cooperation*, London: Wiley (rev. ed. *The Possibility of Cooperation*, Cambridge: Cambridge University Press 1987).

Vieth, M. (2009), *Commitments and Reciprocity. Experimental Studies on Obligation, Indignation, and Self-Consistency*, PhD thesis, Utrecht University.

Voss, T. (1982), "Rational Actors and Social Institutions: The Case of the Organic Emergence of Norms", in: W. Raub (ed.), *Theoretical Models and Empirical Analyses. Contributions to the Explanation of Individual Actions and Collective Phenomena*, Utrecht: ESP, 76–100.

— (1985), *Rationale Akteure und soziale Institutionen*, München: Oldenbourg.

Weesie, J., V. Buskens and W. Raub (1998), "The Management of Trust Relations via Institutional and Structural Embeddedness", in P. Doreian and T. Fararo (eds.), *The Problem of Solidarity: Theories and Models*, Amsterdam: Gordon and Breach, 113–138.

— and W. Raub (1996), "Private Ordering: A Comparative Institutional Analysis of Hostage Games", *Journal of Mathematical Sociology* 21, 201–240.

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

Wippler, R. and S. Lindenberg (1987), "Collective Phenomena and Rational Choice", in: J. C. Alexander et al. (eds.), *The Micro-Macro Link*, Berkeley: University of California Press, 135–152.