Contracts and Promises - An Approach to Pre-play Agreements[*]

Topi Miettinen[†]

SSE/EFI Working Paper Series in Economics and Finance No 707, July 2008

Abstract

In line with the widely applied principle of just deserts, we assume that the severity of the penalty on a contract offender increases in the harm on the other. When this principle holds, the influence of the efficiency of the agreement on the incentives to abide by it crucially depends on whether actions are strategic complements or substitutes. With strategic substitutes, there is a conflict between Pareto-efficiency and the incentives to abide. The opposite tends to be true when actions are strategic complements. The results are interpreted in the context of legal contracts and in that of informal mutual promises.

JEL Classification C72, C78, K12, Z13

KEYWORDS: partnerships, contracts, pre-play communication, legal enforcement, social norms, guilt

[*] This paper is based on chapter 2 of my Ph.D. thesis at University College London. Part of the reasearch was carried out during stay at the Max Planck Institute of Economics, Jena. I am grateful to the Yrjö Jahnsson Foundation for financial support. Thanks to Daniel Friedman, Steffen Huck, Philippe Jehiel, Martin Dufwenberg, David Myatt, Birendra K. Rai, Joel Sobel, Christoph Vanberg, and Anthony Ziegelmeyer for comments.

[†] *Affiliation:* SITE, Stockholm School of Economics. *Address:* Sveavgen 65, PO Box 6501, SE-113 83 Stockholm, Sweden *E-mail:* topi.miettinen@hhs.com

1 Introduction

In societies across the globe, and throughout centuries, there has been a wide public consensus that "punishment should fit the crime" in the sense that more severe crimes which harm others more should be more severely punished. Hamilton and Rytina (1980) and Carlsmith et al. (2002) confirm this consensus in a sociological and in a social psychological study, respectively.[1]

Not only is there societal agreement on this principle of just deserts, but also legal codes, since the code of Hammurabi, largely abide by it. The idea of compensating the plaintiff is central in contract law in particular. Restatement (Second) of Contracts, one of the most well-recognized and frequently-cited legal treatises in Anglo-American jurisprudence, states the following[2]:

"The traditional goal of the law of contract remedies has not been compulsion of the promisor to perform his promise but compensation of the promisee for the loss resulting from breach. ... In general, therefore, a party may find it advantageous to refuse to perform a contract if he will still have a net gain after he has fully compensated the injured party for the resulting loss."[3]

Outside the legal contractual domain, there is evidence that people are intrinsically motivated to abide by informal agreements, and that these private preferences for keeping and breaching promises reflect the just deserts principle. Gneezy's (2005) and Sutter's (2008) findings suggest that people trade off the benefits of lying against the harm that lying inflicts on the opponent. Given the benefit, when harm is more severe, more subjects prefer not lying.[4] This view is also supported by social psychologists. Hoffman (1982), for instance, suggests that guilt has its roots in a distress response to the suffering of others which reflects internalized social norms.

Intrinsic and extrinsic motivation to abide by agreements play a crucial role in the enforcement of obligations in partnerships. In a partnership, two parties decide upon a joint strategy which each partner prefers to acting on her own. Lack of enforcement in partnerships often leads to inefficient inputs or withers the prospect of joining forces entirely.

[1] The latter even illustrate that people have a revealed preference for the just deserts motive despite stating a deterrence motive when asked.

[2] In chapter 16. See also Uniform Commercial Code 1-305 cmt. 1.

[3] 'Just deserts' motive has become predominant even in criminal sentencing not only in Britain and in some states in the United states but also in many European countries (Tonry 2001; von Hirsch, 2007).

[4] Breaching a promise is simply a lie about one's future intentions.

An agreement, whether formal and enforceable in court or informal and enforced by social and psychological forces, typically specifies efficiency-improving standards. It often also specifies the consequences of violations of those standards. Both legal, social, and psychological enforcement, in turn, are shaped by justice principles such as that of harm-fitting of compensation and punishments which we will call the principle of "just deserts" hereafter.

The analysis in this paper applies to any formal or informal partnership when the enforcement scheme satisfies the principle of just deserts in the above-stated harm-fitting sense. The main question studied in this paper asks, how does enforcement according to the principle of just deserts influence efficiency in partnerships. Should enforcement take into account the specific nature of the strategic environment, and if so, how?

It is shown that an enforcement scheme satisfying the principle of just deserts does pretty well in partnerships where inputs are strategic complements.[5] As contract efficiency is improved, the harm inflicted by a marginal contract violation increases, and thus a marginal penalty, which increases in the harm, is stronger if the contract is more efficient. Therefore under specific conditions, if enforcement can improve the status quo at all, it can also enforce a Pareto efficient agreement. This is so, even when without any enforcement, the gain of deviating from the mutual agreement increases in the efficiency of that agreement.

On the contrary with strategic substitutes, the enforcement scheme should look like the opposite of just deserts to provide better incentives to abide by more efficient contracts. As a matter of fact, the Pareto efficiency of the contract and the incentives to stick to it are in direct conflict in those games when just deserts hold. The harm inflicted on the other by a marginal contract violation, and therefore the marginal punishment, decreases as the efficiency of the agreement is improved. Moreover, the gain from the marginal violation increases in efficiency. Both these forces go against the efficiency of the contract.

Contrary to the very economic motivation of enforcement, the just deserts principle provides the weakest incentives for the most efficient contracts when inputs are strategic substitutes. To promote efficiency, punishments should rather be inversely related to the harm on other, or at least not depend on the harm. Yet, this recommendation is in sharp contrast with our basic intuitions of justice and thus it poses a challenge to the

[5] Bulow et al. (1985) introduce and define the concepts of strategic complements and strategic substitutes. Actions are strategic complements if the incentive to increase one's action increases in the action of the other. The opposite holds with strategic substitutes.

design of contracts and their enforcement.

Our results bear implications to two strands of literature. First, Becker (1968) and the subsequent literature[6] on non-strategic 'markets' of criminal activity point out that the just deserts principle is reflected in optimal legal enforcement designed by a social planner who maximizes the sum of expected utilities. Although it is well understood why just deserts may be *implied by* optimal enforcement in such *non-strategic* markets, surprisingly little is known about the *implications of* just deserts on particular *strategic* microstructures of the economy, such as partnerships. Our result, pointing out the crucial importance of strategic complementarity to the efficiency of the contract, constitutes the first steps to fill in this gap.

Second, building upon Farrell (1987, 1988), there is a literature on pre-play communication of intentions.[7] The current paper extends this literature by allowing deviations from pre-play messages or agreements to be costly.[8] These costs are assumed to satisfy the just deserts principle. The cost could be driven by unmodelled social pressure and social punishments carried out by the victim or outsiders.[9] Alternatively, breaching may trigger an emotional reaction, such as guilt or shame, in the offender. If the offender has internalized the just deserts principle, then the negative valence of the emotion increases in the harm inflicted on the other.[10]

Our model helps to understand experimental patterns in public good frameworks. Communication is known to increase cooperation in public good games (Ledyard, 1995), but Suetens (2005) finds that, in the long run, communication induces non-equilibrium levels of cooperation only when actions are strategic complements not when they are strategic substitutes. With complements, cooperation remains higher

[6] Polinsky and Shavell (2000) review the theoretical literature on the public enforcement of law.

[7] See Farrell and Rabin (1996) for an overview. Cheap talk on private information was first analyzed by Crawford and Sobel (1982). In our model, information is complete and information transmission plays no role.

[8] See Demichelis and Weibull (2008) for a recent evolutionary model where players prefer not deviating from pre-play agreements but where this preference is of lexicographically secondary importance. Crawford's (2003) model of boundedly rational pre-play communication assumes that some players always prefer sticking to their pre-play promises.

[9] Darley et al. (2000) illustrate that experimental outsiders' willingness to punish increases in the seriousness of the crime but not in the probability of the offender committing the crime.

[10] Models of social norms assume that people have a preference for abiding by norms, of which promise keeping is just an instance (Bicchieri, 2006; Lopez-Perez, 2008). Given that these preferences largely reflect societal conceptions of justice (Hoffman, 1982), it is of interest extend these approaches to study preferences satisfying the just deserts principle.

than in one-shot equilibrium even after 30 rounds of repetition, whereas with substitutes cooperation decays in a manner that typical for public goods experiments without communication. Similar evidence can be found in Isaac and Walker (1988)[11] and in Kerschbamer et al. (2008). Given our interest in *equilibrium*, our primary interest is to understand and predict long-run and steady state behavior. These long-run patterns are well captured by our model.

The paper is organized as follows. Section 2 presents the model. Section 3 has the main results. In section 4 the interpretation as informal pre-play agreements enforced by intrinsic guilt feelings is discussed. Section 5 concludes.

2 The model

2.1 The underlying game

For the sake of exposition, we use the terminology of legal contracts and enforcement in this section. The alternative interpretation of the contract as an informal pre-play agreement in one-shot games is discussed in section 4. For simplicity and to focus on partnerships, we limit our analysis to two-player games.

The underlying interaction is given by the *underlying game* $\Gamma = \{S_i,\ u_i(s) : S \to R,\ i = 1, 2\}$. The action set of player i in the underlying game is a finite set S_i.[12] A combination of actions is an *action profile* $s = (s_1, s_2) \in S = S_1 \times S_2$. We mainly focus on *finite games with ordered strategies* where higher actions are associated with higher contributions to the partnership. Without loss of generality we label the actions from 0 to n_i, $S_i = \{0, ..., n_i\}$, and we call n_i the *maximal action* of player i.

The *underlying game payoff* to player i is $u_i(s)$. Parties are assumed to be risk neutral; thus $u_i(s)$ corresponds to the monetary compensation of player i. For any given action of player i, her *payoff is increasing in the action of the other player*, j. Reflecting non-increasing marginal productivity of each effort, *payoff is concave (weakly) in own action*

[11] Isaac and Walker (1988) have treatments with constant (weak complements) and decreasing (strict substitutes) returns to scale in a public good provision experiment with communication. Isaac and Walker suggested that the reason for lower cooperation rates with interior group optima might be the difficultly of identifying and agreeing on an interior group optimum. Our model on informal agreements proposes an alterative, perhaps complementary explanation, which also holds in contexts where, even under strategic complements, the efficient profile is interior.

[12] Results would hold with infinite action sets if payoffs are twice continuously differentiable and the right hand derivative of the penalty is strictly positive at zero.

and in that of the other player, where δ_i and σ_i denote the second differences of player i in own and opponent action, respectively.[13] For simplicity, δ_i and σ_i are constant.

The assumption on monotonicity in opponent action is made without a loss of generality. If the payoffs are decreasing in the opponent's action, we can restore our first assumption by reversing the ordering of each strategy set. This has no effect on the second differences. Thus, games with decreasing payoffs in the opponent's action can be analyzed using the same artillery.

Denote the underlying game best-reply correspondence of player i by $BR_i(s_j)$. We restrict our attention to games with pure strategy Nash equilibria and rule out mixed strategies.[14]

2.2 Agreement

Before the game is played the players can enter into an agreement. An agreement m specifies the actions that the parties have agreed to take. Thus, if $m \in M = S$ is the agreement, then m_1 and m_2 are the *agreed actions* of players one and two respectively. If only player i deviates from the agreement and plays a feasible action $s_i \neq m_i$ then

- the *harm* to j is $h_j : M \times S_i \longrightarrow \mathbb{R}$, $h_j(m, s_i) = u_j(m) - u_j(m_j, s_i)$, and the marginal harm to j is $\eta_j(m_j, m_i) = h_j(m_j, m_i, m_i - 1)$.
- the *gain from breaching* to i is $g_i : M \times S_i \longrightarrow \mathbb{R}$, $g_i(m, s_i) = u_i(s_i, m_j) - u_i(m)$ and the marginal gain is $\gamma_i(m_i, m_j) = g_i(m_i, m_j, m_i - 1)$.

Notice that the primary interest is on downward deviations since upward deviations never harm the opponent. Thus, marginal harm and gain are defined in terms of a downward deviation.

To simplify exposition, we adopt the following concepts. For $m \in S$ and for $k \in \mathbb{Z}$, let us call $m + k = (m_1 + k, m_2 + k)$ a symmetric change of actions by k vis-à-vis m.

[13] For all s,

$$\delta_i = u_i(s_i + 1, s_j) - u_i(s_i, s_j) - [u_i(s_i, s_j) - u_i(s_i - 1, s_j)] \leq 0.$$

and

$$\sigma_i = u_i(s_i, s_j + 1) - u_i(s_i, s_j) - [u_i(s_i, s_j) - u_i(s_i, s_j - 1)] \leq 0.$$

[14] An extension to mixed strategies could be easily done. However, the enforcement of mixed strategy agreements is questionable since randomized choices are not verifiable (Abreu, 1988). Thus, perhaps a more natural extension is towards infinite action sets (see the previous footnote). We could also allow for agreements that condition the agreed actions on outcomes of pre-game joint lotteries (Aumann, 1974).

Notice that m itself does not have to be a symmetric action profile as long as both actions are increased or decreased by the same amount. For $k \in \mathbb{Z}$, the effect on the marginal benefit, on the marginal harm, and on the agreed payoff due to such a change are thus $\gamma_i(m+k) = \gamma_i(m_i+k, m_j+k)$, $\eta_i(m+k) = \eta_i(m_i+k, m_j+k)$ and $u_i(m+k) = u_i(m_i+k, m_j+k)$, respectively. These latter two concepts allow us to study the incentive and efficiency effects of symmetric changes of agreements. Yet, we will not restrict our attention to such changes only.

2.3 Enforcement

If partners have agreed on m, a party pays a penalty or a compensation if she deviates unilaterally and her deviation is detected and enforceable in court. Whether the payment is a penalty paid to a third party or a compensation to the other party, does not play an important role since our main focus is on the incentive to respect or breach an agreement which does not depend on the recepient of the compensation.

The magnitude of the compensation depends on the harm inflicted on the other and is given by the function $f : \mathbb{R} \to \mathbb{R}^+$. The goal of the paper is to investigate the implications of the class of compensation functions which are increasing and convex (weakly) in harm. Moreover, if breaching the contract does not harm or it benefits the opponent, the player is not punished. Nonetheless, compensation is strictly positive if the inflicted harm is strictly positive.[15]

Our assumptions allow for a number of possible compensation functions. An example of a function with all the assumed properties is

$$f(h) = \max\{h, 0\}^{\varphi}, \tag{1}$$

where $\varphi \geq 1$. Setting $\varphi = 1$ gives us the exact compensation or crime-fitting applied in the Hammurabi code, incorporating the principles such as "eye for an eye and tooth for a tooth". A fixed punishment

$$f(h) = \begin{cases} \gamma, \text{ if } h > 0 \\ 0, \text{ otherwise} \end{cases}, \tag{2}$$

[15] First, $f(h) = 0$, if $h \leq 0$, and $f(\tilde{h}) \geq f(\hat{h})$ if $\tilde{h} > \hat{h} > 0$. Second, for any $h, \tilde{h}, \hat{h} \in \mathbb{R}$, $f(h) \geq \lambda f(\tilde{h}) + (1-\lambda) f(\hat{h})$ $\forall \lambda \in [0,1]$ if $h = \lambda\tilde{h} + (1-\lambda)\hat{h}$. Weak convexity ensures that marginal arguments suffice when studying whether contracts will be abided by. This implies that the model can capture some experimental findings that a model with a concave compensation function can not (see section 5). Limited liability and the implied non-convexity is briefly discussed in the concluding section.

is not allowed for, however, since this function is concave in harm (because of the discontinuity at the origin).

2.4 Incentive compatible agreements

Suppose the players have agreed to play m, and the exogenously given probability of detecting an enforceable deviation from an agreement is $\theta \in [0,1]$. When the model is interpreted as one with intrinsic enforcement by emotions of guilt guilt, the parameter $\theta_i \in [0,\infty)$ is individual-specific and captures the proneness to guilt. Naturally neither the intrinsic and the extrinsic punishment nor the two interpretations of the parameter θ shold be considered commeasurable.

Let $\Gamma(m;\theta)$ denote the non-cooperative interaction following the agreement m. The net payoff of player i who has agreed on an enforceable contract m can be written as $U_i : M \times S \times \Theta \longrightarrow \mathbb{R}$, where

$$U_i(m,s) = \begin{cases} u_i(m) + g_i(m,s_i) - \theta f(h_j(m,s_i)), \text{ if } s_i \neq m_i \ , \ s_j = m_j \\ u_i(s), \text{ otherwise.} \end{cases} \tag{3}$$

The payoff to player j who complies to the agreement and is not compensated for can be written as $U_j(m_j, m_i, m_j, s_i) = u_j(m) - h_j(m, s_i)$.

Note that there is no punishment if both parties deviate from the agreement. It is a natural assumption since it is not illegal for a victim of a crime to defend herself. Contract law even requires the injured party to take measures to avoid losses, which translates into best-responding to breach: "...the injured party is expected to take reasonable steps to avoid further loss. Where he does this by discontinuing his own performance, he avoids incurring additional costs of performance." (Restatement (Second) of Contracts, ch. 347).

Let us define player i's *incentive to breach* an agreement m as the difference between the gain from breaching and the (expected) compensation, given that the other player does not deviate. Once entered into, an agreement will be *abided* by, if each player's incentive to breach is non-positive, assuming that the other player does not deviate - the agreed action of i must be a best reply to the agreed action of j. An agreement is *incentive compatible* for i if

$$B_i(m, s_i; \theta) \equiv g_i(m, s_i) - \theta f(h_j(m, s_i)) \leq 0, \text{ for all } s_i \in S_i. \tag{IC_i}$$

The agreement m is a Nash equilibrium of the transformed game $\Gamma(m;\theta)$ when the incentive compatibility condition holds for both players. Punishments can only strengthen the incentives to play a given profile of actions. Thus, the following holds.

Proposition 1 *If the agreement m_i is a Nash equilibrium of the underlying game, then it is incentive compatible for any θ in [0,1]. If $\theta = 0$, then only Nash equilibria are incentive compatible.*

Notice that, since the opponent's payoff is increasing in one's action, the marginal harm, $\eta_j(m)$, is always positive. Marginal gain from breaching, $\gamma_i(m)$, can be positive or negative since monotonicity in own action is not assumed.[16] Notice moreover that player i will not be punished if she makes both better off by deviating. Consequently, monotonicity even in own action must hold (non-increasing)[17] at any agreement which is *incentive compatible*. We denote this feasible set[18] of agreements by M^F.

Let us first show that, within this set, non-positive marginal incentive to breach is necessary and sufficient for incentive compatibility. To simply formulate a marginal incentive condition, we define *the marginal incentive to breach*, $\beta_i(m,\theta) = \gamma_i(m) - \theta f(\eta_j(m))$. Clearly, $\beta_i(m,\theta)$ characterizes player i's marginal breaching incentive in M_i^F.

Proposition 2 *Let m_i not be a best-reply to m_j in the underlying game, let $m_i \in M_i^F$ and let m_i differ from the maximal and the minimal action, i.e. $m_i \notin \{0, n\}$. Then an agreement m is incentive compatible if and only if the marginal incentive to breach is non-positive, $\beta_i(m,\theta_i) \leq 0$.*

The simple technicalities behind this result are as follows. The fact that the payoff is concave in the opponent's action implies that the harm h_j is a convex function of s_i. This is because the harm is just the negative of the underlying game payoff of j as a function of s_i given m_j. The negative of the payoff of j is rescaled by adding $u_i(m)$ to all payoffs, i.e. $h_j(m, s_i) = u_j(m) - u_j(m_j, s_i)$. Thus, by the assumption that the punishment is convex in harm, the punishment is convex in s_i as a composite of two convex functions. On the other hand, the underlying game payoff u_i is concave in s_i and, therefore, also the gain from breaching, $g_i(m, s_i)$, is concave. Consequently, checking that neither prefers breaching the agreement marginally is necessary and sufficient for an agreement to be incentive compatible.[19]

[16] Remember that $\eta_j(m)$ and $\gamma_i(m)$ are defined as effects due to a marginal *downward* deviation.

[17] Except for $m_i = n_i$ of course.

[18]
$$M^F = \cap_{i=1,2} M_i^F \text{ where } M_i^F = \{m | \gamma_i(m_i + 1, m_j) \leq 0\}. \tag{4}$$

[19] Notice that it is crucial that we assume that f is a weakly convex function. Otherwise non-marginal

3 Analysis

3.1 Strategic substitutes

By definition, actions are strategic substitutes if i's incentive to reduce her action increases in the opponent's action. Formally, for all s, $u_i(s_i, s_j) - u_i(s_i, s_j - 1) - [u_i(s_i - 1, s_j) - u_i(s_i - 1, s_j - 1)] = \phi_i \leq 0$. Let us now show that as the Pareto efficiency of the agreement is improved vis-à-vis the underlying game equilibrium status quo, the incentives to abide by it are weakened.

When payoff is increasing in the opponent's action and one wishes to strike an agreement which improves efficiency, both must agree to increase their investment in the partnership.[20] To understand the effects of increasing agreed contributions, we will study the effect of increasing own contribution and that of the partner, each at a time.

Each party dislikes unilateral increments of her own contribution above the equilibrium since, within the feasible set of agreements M_i^F, this necessarily reduces payoff. Moreover, since payoffs are concave, a party's marginal incentive to breach, $\gamma_i(m)$, increases if her own agreed action is increased. On the other hand, her partner likes such unilateral increases of the party's contribution since the former's payoff is increasing in the action of the latter. Nonetheless, these marginal payoff-increases gradually decline since the payoffs are concave. Thus, the the marginal harm on the partner due to a marginal breach of the agreement is decreasing. As a combination of these two effects, changing a party's own agreed action unambiguously increases the incentive to breach.

On the other hand, due to actions being strategic substitutes, a party's marginal downward deviation pays off better if her partner's agreed action is higher; moreover in this case, less harm is inflicted on the partner. Thus, increasing j's agreed action also unambiguously increases i's marginal incentive to breach. The following lemma summarizes.

deviations might pay off although a marginal deviation does not.

[20] This is due to equilibrium actions being best replies to each other and to the fact that payoff is increasing in the opponent's action.

Lemma 1

$$\begin{aligned}\gamma_i(m_i+1, m_j) - \gamma_i(m_i, m_j) &= -\delta_i \\ \eta_j(m_i+1, m_j) - \eta_j(m_i, m_j) &= \sigma_j \\ \gamma_i(m_i, m_j+1) - \gamma_i(m_i, m_j) &= -\phi_i \\ \eta_j(m_i, m_j+1) - \eta_j(m_i, m_j) &= \phi_j\end{aligned}$$

With strategic substitutes, $\phi_i < 0$, the effects in lemma 1 on the marginal gain γ_i are positive and the effects on the marginal harm are negative. Since an efficiency improving contract must specify actions being increased from the equilibrium status quo, a conflict is implied between the Pareto efficiency of an agreement and the incentives to stick to it in games with strategic substitutes.

Theorem 1 *Let actions be strategic substitutes. If efficiency is improved vis-à-vis an interior equilibrium, s^*, then the marginal gain from breaching is increased and the marginal harm is decreased.*

Alternatively, let s be an agreement which is more efficient than s^*, and let θ be such that a party is indifferent between keeping and breaching s. Then the party will breach all agreements more efficient than s and for all probabilities larger than θ, the party will keep all agreements less efficient than s and more efficient than s^*.

3.2 Strategic complements

Let us now turn to games where actions are strategic complements. By definition, actions are strategic complements if $\phi_i \geq 0$. It is easily seen that strategic complementarity does not change the impact of own agreed action on incentives to abide by the contract. Yet, the effect of increasing the *opponent's* agreed action is now the opposite.

Increasing efficiency requires upward adjustments in both agreed actions at the same time. Whereas with strategic substitutes the opponent action effect goes hand in hand with the own action effect thus deteriorating incentives as efficiency is improved, with strategic complements, the opponent-action effect downplays the anti-efficiency impact of the own agreed action. It is not clear a priori whether the effect of the own action dominates the effect of the opponent action on the marginal incentive to breach. If it does, then the marginal incentive to breach again comes into opposition with Pareto efficiency.

To study the issue in more detail, we can alternatively decompose the effect on incentives into the benefit-effect (lines 1 and 3 in lemma 1) and the harm effect (lines 2 and 4 in lemma 1). When $\phi_i + \delta_i \geq 0$, the marginal gain from breaching is decreasing as efficiency is improved. In this case, strategic complementarity is very strong - so strong that, if $\phi_i + \delta_i \geq 0$ holds for both parties, maximal actions are efficient and, moreover, they constitute a Nash equilibrium of the underlying game. Therefore, by proposition 1, the maximal actions are incentive compatible.[21] The case of strong complementarity is somewhat uninteresting for us since enforcement plays no role in achieving efficiency. Agreements then enact the part of a mere coordination device or a convention when choosing among multiple equilibria. Our theorem below establishes that, even in the more interesting case where actions are weaker strategic complements and the efficient profile is not an equilibrium, contracts may achieve first-best efficiency if any improvements to efficiency can be achieved at all. It considers improving efficiency through symmetric changes vis-à-vis the best status quo, i.e. underlying game equilibrium, (according to Pareto ranking).

Theorem 2 *Let s^* be the most efficient interior underlying game equilibrium. Let $\phi_j + \sigma_j \geq 0$. Let $s^* + k$ be is more efficient than s^* and incentive compatible for i and let $s^* + k - 1$ be less efficient than $s^* + k$ and not incentive compatible for i. Then an efficient agreement is incentive compatible for i.*

Alternatively, let $s^* + k$ be a symmetric change of actions that makes the agreement more efficient than s^*, and let θ be such that a party is indifferent between keeping and breaching $s^* + k$ and prefers breaching $s^* + k - 1$. Then the party will breach all agreements less efficient than $s^* + k$ and, for all probabilities larger than θ, she will keep all agreements $s^* + K$ more efficient than $s^* + k$.

When strategic complementarity is weaker, the marginal gain from breaching increases as efficiency is improved, $\delta_i + \phi_i < 0$. If complementarity is so weak that it does not even dominate the concavity effect in the opponent action (let us call this *the opponent-concavity effect*) $\sigma_j + \phi_j < 0$, then also the marginal harm decreases as the agreement is made symmetrically more efficient. Incentives stand again in contradiction with Pareto efficiency as was the case with strategic substitutes.

Yet, if strategic complementarity is strong enough to downplay the opponent-concavity effect, $\phi_i + \sigma_i \geq 0$, then the marginal harm is non-decreasing and acts as an opposing force to the marginal benefit as efficiency is improved.[22] If the mitigating role played by the increasing harm becomes sufficiently important to eventually level off the increasing breaching incentive, then all agreements that are more efficient than this threshold

[21] Milgrom and Roberts (1990) show that, with strategic complements, equilibria are pareto-ranked with highest contributions equilibrium having the highest rank.

[22] Its effect is strengthened, if the punishment function is strictly convex.

agreement are incentive compatible. The discord between efficiency and incentives may thus be circumvented if the actions are strategic complements.

4 Informal agreements and guilt

Let us now pursue the intrinsically motivated approach. The intrinsic approach associates the function f with guilt feelings, or disutility, about breaking the promise, m. In this intrinsic context, the parameter θ is an individual-specific one measuring sensitivity or proneness to guilt, $\theta_i \in [0, \infty)$. Probability of detection plays no role since the punishment is entirely intrinsic and a deliberate transgressor knows that she has transgressed the agreement. The assumption that informal promises about intentions are not merely cheap talk, and that people may feel guilty about breaching promises, is very intuititive. There is also abundance of indirect (Sally, 1995) and some more direct evidence (Vanberg, 2008) in favor of this view.

Notice that, if both players had zero proneness to guilt, the model presented above coupled with a communication protocol would reduce to the renowned cheap talk model of Farrell (1987). Our model thus extends the chap talk on intentions by Farrell. As in Farrell, implicit in our formulation is an assumption that players have a common understanding of what they agree upon, the mapping from agreements to prescribed actions is exogenously given and common knowledge.[23]

Guilt has been discussed in several papers since Akerlof (1980) who develops a model of conformity to social norms, or Frank (1988) who argues that it may well be materially profitable for an agent to have a conscience - a dislike for disobeying social norms.[24] Kandel and Lazear (1992) study a model guilt and shame in partnership situations like ours, but not addressing the issue of strategic complementarity and comparative enforcement power of agreements. More specific and experimentally motivated models are proposed by Ellingsen and Johanneson (2004), Bicchieri (2006), Lopez-Perez (2008), and Dufwenberg and Battigalli (2007).[25] The former three are more traditional outcome-based models where the agreement is considered as part of the outcome or it is an exogenously given social norm. The present model extends

[23] This is in opposition to approaches where the meaning of messages is derived as part of the equilibrium leading to multiple equilibria due to multiplicity of interpretations. As opposed to Farrell, the model in this paper does not explicitly model the communication.

[24] Rotemberg (1994), and Bester and Güth (1999) point out the importance of strategic complementarity in rationalizing the choice to become altruistic.

[25] See also Kaplow and Shavell (2007) for a non-strategic model of guilt.

the existing outcome-based models of guilt by allowing guilt to be increasing in the inflicted harm. This feature is crucial for our results.

Dufwenberg and Battigalli (2007) model guilt as an explicit belief-dependent motivation, aversion for letting down the expectations of the other. It thus falls into the category of psychological games (Geanokoplos et al., 1989). While there is experimental evidence that in some contexts preferences are belief-dependent, there is also evidence that the agreements per se matter (Vanberg, 2008). The advantage of outcome-based modelling is its simplicity, amenability to revealed preference interpretations (Cox et al., 2007) and weaker reliance on the equilibrium assumption. Yet, our model has a straightforward psychological game interpretation studying whether a given profile m can be sustained as a psychological Nash equilibrium[26]. In that interpretation the agreement, as an argument of the utility function, is interpreted as the equilibrium profile of beliefs and one is interested whether any player has an incentive to deviate from that profile. If not, we have a psychological Nash equilibrium.

The results of the previous sections fully carry over to the alternative interpretation of intrinsic enforcement of infomal promises. Moreover, the informal model can be extended to allow guilt to respond reciprocally to generosity. In this extension, guilt is allowed to depend not only on the harm inflicted on the partner but also to increase in the payoff the player would receive if both kept their part of the agreement, the player's agreed payoff. The more the partner agrees to give, the more guilty a player may feel about letting the partner down. This approach allows us to prove a result analogous to theorem 2. We can also derive a corollary which strengthens the result: if anything more efficient than a unique UG equilibrium is incentive compatible for i, then an efficient agreement is also incentive compatible.

We will now proceed with this extension of the model. To introduce the generosity-effect formally, we adopt a measure for the generosity of the agreement. Let *the lowest Nash payoff of player i* be defined as

$$u_i^* = \min_{s \in NE(\Gamma)} u_i(s)$$

where $NE(\Gamma)$ is the set of pure Nash equilibria of the underlying game. The vector of lowest Nash payoffs is $u^* = (u_1^*, u_2^*)$. For player i, the lowest Nash payoff is the worst case scenario if the there is no agreement in place.[27] The *agreed payoff* mea-

[26] See Geanokoplos et al. (1989) for a definition and discussion.

[27] u_1^* and u_2^* can result from different action profiles.

sures generosity by indicating how much more than u_i^* the player gets[28] if both respect the agreement, $v_i(m) = u_i(m) - u_i^*$. Player i's *guilt cost,* $\theta_i f(v_i(m), h_j(m, s_i))$ would now depend on this agreed payoff in addition to the inflicted harm. Notice yet that even partners who were infinitely prone to guilt could not agree on all agreements in all circumstances since each party would have no trouble making efficiency-improving unilateral deviations.

Guilt is assumed *weakly increasing* in the agreed payoff, v_i, and in the harm, h_j, and whenever both are strictly positive guilt is strictly positive, otherwise not. If the player inflicts no harm on the opponent or if the agreement treats the player ungenerously (the agreed payoff is equal or below the worst Nash payoff), then the player will not feel guilty about breaking a promise,

$$\begin{array}{ll} f(v_i, h_j) > 0, & \text{if } h_j > 0, v_i > 0 \\ f(v_i, h_j) = 0, & \text{if } h_j \leq 0 \text{ or } v_i \leq 0 \end{array}. \tag{5}$$

An example of a suitable guilt function is

$$f(v_i(m), h_j(m, s_i)) = \max\{v_i(m), 0\}^\gamma \max\{h_j(m, s_i), 0\}^\varphi. \tag{6}$$

The entire game preferences of this form with $\gamma = \varphi = 1$ are closely related to tractable preferences of Cox et al. (2006), and thus, our model can be broadly considered as a tractable model of guilt.[29]

When accounting for the additional properties of guilt, we can show that a version of theorem 2 holds.

Theorem 3 *Let s^* be the most efficient interior UG equilibrium. Let $u_i(s + k)$ be convex in k and let f be weakly convex in v_i and supermodular[30] in its arguments. Let s^* be the most*

[28] Any reference payoff greater than this worst Nash payoff will do as well. Rabin (1994) derived the worst pareto-efficient Nash equilibrium payoff as the lower payoff bound for a player engaging in cheap talk, for instance.

[29] These latter formalize the following other-regarding preferences: the payoff of i is $(\pi_i^\alpha + \omega\pi_j^\alpha)/\alpha$ where π_i is player's own material payoff, π_j is that of the other player, $\alpha \in (-\infty, 0) \cup (0, 1]$ is an elasticity parameter, and ω is a function capturing reciprocity and other emotional state motivations. Setting $\alpha = 1$, $\omega(m) = \theta_i \max\{v_i(m), 0\}$ and normalizing $\pi_j = u_j(m_j, s) - u_i(m)$ returns our entire game preferences with the above presented guilt cost and $\gamma = \varphi = 1$ as a special case of the preferences in Cox et al. (2006). Of course the truncation, $\max\{h_j(m, s_i), 0\}$, is particular for our model of prescriptive informal norms and it does not arise when modelling equity motivations as in Cox et al. (2006). The relationship to outcome-based fairness models is discussed in the conclusion.

[30] Increasing harm weakly increases the marginal effect of the agreed payoff and vice versa.

efficient interior underlying game equilibrium. Let $\phi_j + \sigma_j \geq 0$. Let $s^* + k$ be is more efficient than s^* and incentive compatible for i and let $s^* + k - 1$ be less efficient than $s^* + k$ and not incentive compatible for i. Then an efficient agreement is incentive compatible for i.

In this framework of informal agreements, we can prove even a more powerful result stated in corollary 1. If there exists an incentive compatible informal agreement that improves efficiency vis-à-vis a status quo with a unique equilibrium, then a Pareto efficient agreement is incentive compatible.

Corollary 1 *Let s^* be the unique underlying game equilibrium. Let $\gamma_i(s^*) = 0$ for $i = 1, 2$. Let $u_i(s + k)$ be convex in k and f be weakly convex in v_i and supermodular in its arguments. Let $\phi_j + \sigma_j \geq 0$. Let $s^* + k$ be is more efficient than s^* and incentive compatible for i. Then an efficient agreement is incentive compatible for i.*

This result relies, first, on the fact that an ungenerously treated partner behaves opportunistically, and second, on the fact that the equilibrium generosity is zero. Thus, at the equilibrium the marginal guilt cost is necessarily smaller than the marginal gain from breaching. Thus, if the increasing guilt cost ever levels off the increasing marginal benefit as the Pareto efficiency of the agreement is improved due to a symmetric change of agreed actions vis-à-vis the equilibrium, this shift in balance will hold for every more efficient symmetric agreement due to the convexity properties.

Some assumptions that have been made in Sections 2.3 and 2.4 become particularly compelling in the context of informal agreements. More specifically, if the other breaches the agreement, a victim does not typically feel bad about protecting herself from exploitation. As far as the informal agreement interpretation is concerned, we thus have assumed that, if the other party breaches, then a party feels no guilt whatsoever. This guarantees that marginal violations of the agreement will have non-marginal implications. Bicchieri (2006, ch. 1) discusses in length why conditional conformism captured by this feature is necessary in any reasonable account of social norms.[31] The simple idea hidden in conditional conformism when applied to partnerships is that people do not like to be exploited suckers.

5 Discussion

The model in this paper illustrates the implications of "punishments (or compensation) that fit the harm" on the enforcement of bilateral partnerships. We show that,

[31] See also Lopez-Perez (2008).

in partnerships where inputs are strategic substitutes, efficiency and the incentives to provide input are in conflict. Nevertheless, partnerships with strategic complements may avoid such conflicts, if strategic complementarity is sufficiently strong. If there is an efficiency improving symmetric agreement that a player would not abide by and another more efficient agreement that she would abide by, then she is willing to abide by a *first best* agreement.

In section 4, we studied informal pre-play agreements in one-shot games which are enforced by intrinsic motivation to avoid guilt about breaching. The pre-play negotiation protocol was unmodelled, however. Thus, it does not matter for the results how the contract is agreed upon. The conclusions would not be altered if the agreement was established in a commonly known code of conduct - a social norm - rather than in negotiations. Thus, one can alternatively use the model to analyze the comparative enforcement power of social norms in various free-riding and public good provision contexts. Yet, this would generally require an extension to a multiplayer setting involving modelling choices on how punishments depend on the profile of harms, for instance. This latter is essentially an empirical question and I am not aware of convincing evidence favoring any given modelling approach. Thus, the multiplayer extension is perhaps better left for future research.

One application of interest are symmetric norms in symmetric games. Such norms can be essentially considered as equity norms. Not surprisingly, when interpreted as a model of social norms, our model shares some features with inequity aversion models (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). An inequity averse offender also feels guilty about transgressing a norm if it generates advantageous inequality. Thus, as long as we consider symmetric agreements in symmetric games only, the punishment can be considered as a disutility for violating an equity norm. Our results hold for this particular application as well: if the actions are in material terms strategic substitutes, inequity aversion provides weaker incentives for efficient play than for inefficient play, for instance.

In the literature on infinitely repeated games, grim strategies require offenders of (Nash) equilibrium play to be punished forever. These strategies provide the strongest enforcement power, and in fact, any pre-play agreement on a stationary play can be enforced if each party gets more than her reservation payoff in each period.[32] Neverthe-

[32] Any agreement m on a stationary play giving each player more than $(1-\delta_i)BR_i(m_j)+\delta_i\underline{u}_i$, where δ_i is i's discount factor and $\underline{u}_i$ is player i's reservation payoff, can be sustained in a Nash equilibrium of an infinitely repeated game. Each player must receive more than a mark-up above the reservation payoff, if she is impatient. Abreu (1988) illustrates that in a *perfect* equilibrium the most severe punishment

less, if the severity of the punishment (time spent punishing) is an increasing function of the harm inflicted, as required by just deserts, a violator who inflicts minor harm cannot be very severely punished. Thus, if the stage game actions are strategic substitutes, efficiency of a stationary pre-play agreement and the incentives to abide by it are inversely related. The most efficient stationary agreements are the prime suspects not to be enforceable when just deserts hold. The conclusions of the model will continue to hold if the interaction is repeated and punishments cannot be agreed upon but are exogenously given and follow the just deserts principle. Parameter θ_i is now associated with a player's dicount factor and f maps the inflicted harm into a punishment path.

For the extrinsic interpretation of the enforcement model, it is interesting to speculate on how our results would be influenced by introducing liability constraints on the compensation that the offender pays to the plaintiff. Intuitively, this should have no effect on the result on strategic substitutes as drastic underlying game best-reply deviations would be perceived for low levels of efficiency where harm is large and thus compensation hits the limit. Yet, this increased breaching incentive would tend to disappear with efficiency, in line with out established result. Nevertheless, when actions are strategic complements, an ambiguity would appear at high efficiency levels where harm is so large that compensation exceeds the liability constraint. With intermediate strategic complementarity where both the marginal gain and the marginal harm increase in efficiency, limited liability would potentially imply a u-shaped association between efficiency and incentives to breach.

That offenders should receive their just deserts is a widely applied principle in human interaction. It is so natural to us that we hardly pay attention how substantially it influences the formal and informal institutions that surround us and constrain our lives. In strategic rather than non-strategic interaction, it is perhaps even more important to thoroughly understand how justice principles shape the incentives to abide by agreements. This is because only in strategic interaction even marginal violations may induce non-marginal and unpredictable consequences and prevent mutually beneficial joint ventures from flourishing. This paper points out that the strategic nature of the underlying interaction, whether one-shot or repeated, and the institutional features of punishments, whether formal or informal, have crucial and surprising interdependencies that should not be neglected if we are concerned about providing optimal third party enforcement for voluntarily initiated partnerships.

has a stick-and-carrott structure and thus holding the other to her reservation payoff for infinitely long payoff is not feasible.

6 Appendix

Here we give proofs for the extended model where cost is a function of generosity of j, v_i, in addition to the harm (Section 4). Proofs of the case in the text where f is a function of h_j only are mainly special cases and where they are not, it will be indicated.

6.1 Proof of proposition 2

Lemma 2 returns proposition 2 in the text (where f is a function of h_j only) as a special case.

Lemma 2 *Let Γ be finite. Let $m_i \neq \{0, n\}$, $m_i \in M_i^F$ and $m_i \notin BR_i(m_j)$. Then an agreement m is incentive compatible if and only if $\beta_i(m, \theta_i) \leq 0$.*

We will show that (IC_i) does not hold if and only if $\beta_i(m, \theta_i) > 0$. Let $\beta_i(m, \theta_i) > 0$. By the definition of $\beta_i(m, \theta_i)$, $B(m, m_i - 1; \theta_i) = \gamma_i(m) - \theta_i f(v_i(m), \eta_j(m)) > 0$ and thus incentive compatibility, (IC_i), is violated.

Let incentive compatibility, (IC_i), be violated. Thus, there is s_i' such that $B_i(m, s_i'; \theta_i) > 0$. Suppose to the contrary that $\beta_i(m, \theta_i) \leq 0$ and thus

$$u_i(m_i - 1, m_j) - u_i(m_i, m_j) \leq f(v_i(m), h_j(m, m_i - 1)).$$

There are two cases to consider $u_i(m_i - 1, m_j) - u_i(m_i, m_j) < 0$ and $u_i(m_i - 1, m_j) - u_i(m_i, m_j) \geq 0$. In the first case, it is also true that $u_i(m_i, m_j) - u_i(m_i + 1, m_j) < 0$ since otherwise m_i is an underlying best-reply to m_j by the concavity of u_i in its first argument. Now $u_i(m_i, m_j) - u_i(m_i + 1, m_j) < 0$ implies that $m_i \notin M_i^F$ which is a contradiction. In the second subcase $u_i(m_i - 1, m_j) - u_i(m_i, m_j) = \gamma_i(m) > 0$ and thus $f(v_i(m), h_j(m, m_i - 1)) > 0$ since $\beta_i(m, \theta_i) \leq 0$. By assumption, harm increases in deviations further downwards. Also by assumption guilt cost is convex in h_j and u_j is concave in s_i. Thus harm is convex in s_i and the guilt cost is also convex in s_i as a composite of two convex functions. On the other hand by assumption, the payoff u_i is concave in s_i and thus the gain from breaching $u_i(s_i, m_j) - u_i(m_i, m_j)$ is concave in s_i. Thus if $\beta_i(m, \theta_i) \leq 0$ then $B(m, s; \theta_i) \leq 0$ for all $s_i < m_i$. We have a contradiction.

6.2 Proof of lemma 1

$$\begin{aligned} &\gamma_i(m_i + 1, m_j) - \gamma_i(m_i, m_j) \\ &\quad = u_i(m_i, m_j) - u_i(m_i + 1, m_j) - [u_i(m_i - 1, m_j) - u_i(m_i, m_j)] \end{aligned}$$

$$= -\delta_i$$

$$\begin{aligned}&\gamma_i(m_i, m_j+1) - \gamma_i(m_i, m_j)\\ &\quad= u_i(m_i-1, m_j+1) - u_i(m_i, m_j+1) - [u_i(m_i-1, m_j) - u_i(m_i, m_j)]\\ &\quad= -\phi_i\end{aligned}$$

$$\begin{aligned}&\eta_j(m_j, m_i+1) - \eta_j(m_j, m_i)\\ &\quad= u_j(m_j, m_i+1) - u_j(m_j, m_i) - [u_j(m_j, m_i) - u_j(m_j, m_i-1)]\\ &\quad= \sigma_j\end{aligned}$$

$$\begin{aligned}&\eta_j(m_j+1, m_i) - \eta_j(m_j, m_i)\\ &\quad= u_j(m_j+1, m_i) - u_j(m_j+1, m_i-1) - [u_j(m_j, m_i) - u_j(m_j, m_i-1)]\\ &\quad= \phi_j \blacksquare\end{aligned}$$

6.3 Proof of theorem 1

Since s is more efficient than s^*, it must be that $s_i \leq s_i^*$ for $i = 1, 2$. To see this, notice that since the payoff is increasing in the opponent's action and s_i^* is a best-reply to s_j^* , for any $s_i \leq s_i^*$, $u_j(s^*) \geq u_j(s_j, s_i^*) > u_j(s)$. Thus each action at s must necessarily be weakly greater than at s^* to increase efficiency. But in that case, by lemma 1, $\gamma_i(s) > \gamma_i(s^*)$ and $\eta_j(s) < \eta_j(s^*)$. $\blacksquare$

6.4 Proof of theorem 2

Since $s^* + k$ is more efficient than s^*, it must be that $k > 0$ by the same argument as in the proof of 1. Thus for a symmetric change of actions to be efficiency-improving, it must hold that $k > 0$.

Let first $\phi_i + \delta_i < 0$. Since $s^* + k$ is incentive compatible for i and $s^* + k - 1$ is not incentive compatible for i, by lemma 3 (below), for every $K > k$, $s^* + K$ is incentive compatible for i.

Let us show that i's payoff-maximizing profile among symmetric increases of actions, denote it by $s^* + \overline{k}_i$, exists and satisfies $\overline{k}_i \geq k$. To see this, suppose first that $\sigma_i + \delta_i + 2\phi_i \geq 0$. The assumption implies that $u_i(s + K)$ is convex in K. Now, by the argument above, $u_i(s^*) - u_i(s^* - K) > 0$ for any $K > 0$. Therefore, every symmetric increase of actions increases the payoff of i vis a vis s^* and by the boundedness of the strategy space, there exists $\overline{k}_i$ such that $s^* + \overline{k}_i$ maximises i's payoff. Suppose alternatively that $\sigma_i + \delta_i + 2\phi_i < 0$. Then $u_i(s + K)$ is strictly concave in K. By assumption, $u_i(s^* + k) \geq u_i(s^* + k - 1)$. Since the strategy set is finite, a maximizer $s^* + \overline{k}$ among symmetric increases of actions exists and it satisfies $\overline{k} \geq k$. By the same arguments, there is a $\overline{k}_j \geq k$ which maximizes j's payoff among the symmetric changes of actions.

Let $\overline{k}$ be defined as $min(\overline{k}_i, \overline{k}_j)$. Notice that by the argument above, $s^* + \overline{k}$ is incentive compatible for i.

Let us show that $s^* + \overline{k}$ is efficient if $\phi_j + \delta_j < 0$. (We will return further below to the case $\phi_j + \delta_j \geq 0$.) It is easy to see that no symmetric change of actions is Pareto-preferred to $s^* + \overline{k}$. Moreover no unilateral change of an action is Pareto-preferred since s^* is an equilibrium and thus, by lemma 3.1, $\gamma_i(s_i^* + K + 1, s_j^* + K) \geq 0$ for $K \geq 0$ and payoffs are concave in own action. Thus i's payoff is worse if her action alone is increased from $s^* + \overline{k}$. On the other hand, if i's action is decreased, j's payoff decreases by assumption. Combinations of symmetric changes of actions and unilateral changes of one particular action reach any strategy where $s_i > s_i^*$ yielding the result.

If $\phi_j + \delta_j \geq 0$, there may be no symmetric change of actions that is efficient. This is the case if at $s^* + \overline{k}$ as defined above $\gamma_j(s_j^* + \overline{k} + 1, s_i^* + \overline{k}) < 0$, in which case j can improve the payoff of both by deviating upwards. Let thus j's action be increased until her UG best-response is reached (which can satisfy $s_j = n_j$). Since j's agreed action is unilaterally increased, by lemma 3.1, γ_i is weakly smaller and η_j is weakly greater than before the unilateral increase of j's action. Thus, i's marginal incentive to breach is smaller or equal to that at $s^* + \overline{k}$.

If i's marginal gain from breaching is still weakly positive, $\gamma_i(s_i^* + \overline{k} + 1, BR_j(s_i^* + \overline{k})) \geq 0$ and thus, $s_i^* + \overline{k}$ is not i's UG best-reply to none of the actions between $s_j^* + \overline{k}$ and $BR_j(s_i^* + \overline{k})$, then $(s_i^* + \overline{k}, BR_j(s_i^* + \overline{k}))$ is efficient and incentive compatible for i.

If $\gamma_i(s_i^* + \overline{k} + 1, BR_j(s_i^* + \overline{k})) < 0$, then one can increase the agreed action of i until $\gamma_i(s_i^* + \overline{k} + K + 1, BR_j(s_i^* + \overline{k} + K)) \geq 0$. Such K exists since strategy sets are bounded. Since such changes can always be made as a sequence of marginal changes of one of the actions at a time so that $\gamma_i, \gamma_j \leq 0$ (and one with strict inequality), both payoffs and thus efficiency is increased due to these changes. Moreover, in the resulting profile, each action is an UG best-reply to that of the other. Thus, as a Nash equilibrium of the UG, the profile is incentive compatible for i.

Consider now the case, $0 \leq \delta_i + \phi_i$. Now the marginal gain from breaching is decreasing in symmetric changes of actions. At the interior equilibrium, $\gamma_i(s_i^*, s_j^*) \leq 0$ and $\gamma_i(s_i^* + K + 1, s_j^* + K) \geq 0$. Thus for $K \geq 0$, $\gamma_i(s_i^* + K, s_j^* + K) < 0$ and if there is k' such that $\gamma_i(s_i^* + k' + 1, s_j^* + k') \geq 0$, then for every $0 < k < k'$, $\gamma_i(s_i^* + k + 1, s_j^* + k) \geq 0$. Therefore, $s^* + k$ can be incentive compatible but not $s^* + k - 1$ only if $s_i^* + k = n_i$.

If moreover $\gamma_j(s_j^* + k + 1, s_i^* + k) \leq 0$ or if $s_j^* + k = n_j$, then $s^* + k$ is efficient. If $\gamma_j(s_j^* + k + 1, s_i^* + k) > 0$ and $s_j^* + k < n_j$, then one can increase the action of j until

$s_j = BR_j(n_i)$. It is easy to see that this increases efficiency and, moreover, since actions are strategic complements, player i has even more of an incentive to choose $s_i = n_i$. Thus, $(n_i, BR_j(n_i))$ is efficient and incentive compatible for i.

Lemma 3 *Let the underlying game payoff second differences satisfy $\phi_i+\delta_i < 0$ and $\phi_i+\sigma_i \geq 0$. Suppose that $\gamma_i(s-1) \geq \theta_i f(\eta_j(s-1))$. If $\gamma_i(s) \leq \theta_i f(\eta_j(s))$ then $\gamma_i(s+k) \leq \theta_i f(\eta_j(s+k))$ for all $k > 0$.*

By lemma 1, the marginal incentive to breach, $\gamma_i(s+k)$, is increasing and concave in k and the marginal harm on the other, $\eta_j(s+k)$, is non-decreasing and convex in k.

Also $f(\eta_j(s+k))$ is convex and non-decreasing in k since g is convex in η for $\eta_j \geq 0$.

Also since $\gamma_i(s-1) \geq \theta_i f(\eta_j(s-1)) \geq 0$ but $\gamma_i(s) \leq \theta_i f(\eta_j(s))$, we have

$$\begin{aligned} & \gamma_i(s) - \gamma_i(s-1) \\ \leq \;& \theta_i f(\eta_j(s)) - \theta_i f(\eta_j(s-1)). \end{aligned}$$

Thus, since $f(\eta_j(s+k))$ is convex and non-decreasing in k,

$$\begin{aligned} 0 &\leq \gamma_i(s+1) - \gamma_i(s) \\ &= -\delta_i - \phi_i \\ &= \gamma_i(s) - \gamma_i(s-1) \\ &\leq \theta_i f(\eta_j(s)) - \theta_i f(\eta_j(s-1)) \\ &\leq \theta_i f(\eta_j(s+1)) - \theta_i f(\eta_j(s)). \end{aligned}$$

We can proceed by induction to show that for every $s+k$ with $k > 0$, we have $\gamma_i(s+k) - \theta_i f(\eta_j(s+k)) \leq \gamma_i(s) - \theta_i f(\eta_j(s)) \leq 0$. Thus every $s+k$ with $k > 0$ is incentive compatible.

Proof of theorem 3

The proof is along the lines of the proof of theorem 2 above. Yet, instead of applying lemma 3 one should apply 4 below. ■

Proof of the corollary 1

By the arguments of proof of theorem 2 above, there must be $k'_i > 0$ such that $s^*+k'_i$ is incentive compatible for i implying $\gamma_i(s^*+k'_i) \leq \theta_i f(v_i(s^*+k'_i), \eta_j(s^*+k'_i))$. On the other hand s^* satisfies $\gamma_i(s^*) \geq \theta_i f(0, \eta_j(s^*)) = 0$ as the unique UG equilibrium. Thus, the claim follows from theorem 3. ■

6.5 Proof of lemma 4

Lemma 4 *Let the underlying game payoff second differences satisfy $\sigma_j + \phi_j \geq 0$, $\phi_i + \delta_i < 0$ and $2\phi_i + \delta_i + \sigma_i \geq 0$. Let s be symmetric, that is $s_i = s_j$. Let $u_i(s) - u_i(s-1) \geq 0$. Let g be supermodular. Suppose that $\gamma_i(s-1) \geq \theta_i f(v_i(s-1), \eta_j(s-1))$. If $\gamma_i(s) \leq \theta_i f(v_i(s), \eta_j(s))$ then $\gamma_i(s+k) \leq \theta_i f(v_i(s+k), \eta_j(s+k))$ for all $k > 0$.*

By lemma 1, the marginal gain from breaching, $\gamma_i(s+k)$, is increasing and concave in k and the marginal harm, $\eta_j(s+k)$, is non-decreasing and convex in k.

Since $\delta + 2\phi + \sigma \geq 0$ and $u_i(s) - u_i(s-1) \geq 0$, $u(s+k)$ is convex and non-decreasing in k for $k \geq 0$. Thus, $f(v_i(s+k), \eta_j(s))$ is convex and non-decreasing in k since g is convex and non-decreasing in v. Similarly, $f(v_i(s), \eta_j(s+k))$ is convex and non-decreasing in k since g is convex in η for $\eta \geq 0$.

Also since $\gamma_i(s-1) \geq \theta_i f(v_i(s-1), \eta_j(s-1)) \geq 0$ but $\gamma_i(s) \leq \theta_i f(v_i(s), \eta_j(s))$, we have

$$\begin{aligned} & \gamma_i(s) - \gamma_i(s-1) \\ \leq \; & \theta_i f(v_i(s), \eta_j(s)) - \theta_i f(v_i(s-1), \eta_j(s-1)). \end{aligned}$$

Thus, since g is supermodular and convex in its arguments,

$$\begin{aligned} 0 &\leq \gamma_i(s+1) - \gamma_i(s) \\ &= -\delta - \phi \\ &= \gamma_i(s) - \gamma_i(s-1) \\ &\leq \theta_i f(v_i(s), \eta_j(s)) - \theta_i f(v_i(s-1), \eta_j(s-1)) \\ &\leq \theta_i f(v_i(s+1), \eta_j(s+1)) - \theta_i f(v_i(s), \eta_j(s)). \end{aligned}$$

We can proceed by induction to show that for every $s+k$ with $k > 0$, we have $\gamma_i(s+k) - \theta_i f(v_i(s+k), \eta_j(s+k)) \leq \gamma_i(s) - \theta_i f(v_i(s), \eta_j(s)) \leq 0$. Above, we showed that $u_i(s+k) > u_i(s)$ for $k > 0$. Thus every $s+k$ with $k > 0$ is incentive compatible.

References

[1] "The Second Restatement of Contracts.", *Columbia Law Review*, 81, 1

[2] Abreu, D. (1988), 'On the Theory of Infinitely Repeated Games with Discounting', *Econometrica*, 56, 383-396.

[3] Akerlof, G. (1980), 'A Theory of Social Customs, of Which Unemployment May be One Consequence', *Quarterly Journal of Economics*, 94, 749-775.

[4] Aumann, R. J. (1974), 'Subjectivity and Correlation in Randomized Strategies', *Journal of Mathematical Economics*, 1, 67-96.

[5] Becker, G. (1968), 'Crime and Punishment', *Journal of Political Economy,* 76, 169-217.

[6] Bicchieri, C. (2006), *The Grammar of Society*, New York: Cambridge University Press.

[7] Bolton, G., Ockenfels, A., (2000), 'ERC: A Theory of Equity, Reciprocity, and Competition', *American Economic Review,* 90, 166-193.

[8] Bulow, J., Geanakoplos, J., Klemperer, P. (1985), 'Multimarket Oligopoly: Strategic Substitutes and Complements', *Journal of Political Economy*, 93, 488-511.

[9] Carlsmith, K.M., Darley, J.M, Robinson, P.H. (2002), 'Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment', *Journal of Personality and Social Psychology,* 83, 284-299.

[10] Cox, J., Friedman D, Gjerstad S. (2006), 'A Tractable Model of Reciprocity and Fairness', *Games and Economic Behavior*, forthcoming.

[11] Cox, J., Friedman D., Sadiraj, V. (2007), 'Revealed Altruism', *Econometrica*, forthcoming.

[12] Demichelis, S., Weibull, J. (2008), 'Meaning and Games: A Model of Communication, Coordination and Games', *American Economic Review,* forthcoming.

[13] Dufwenberg, M. (2002), 'Marital Investment, Time Consistency & Emotions', *Journal of Economic Behavior and Organization*, 48, 57-69

[14] Ellingsen, T. , Johanneson, M. (2004), 'Promises, Threats, and Fairness', *Economic Journal*, 114, 397-420.

[15] Farrell. J. (1987), 'Cheap Talk, Coordination, and Entry' *Rand Journal of Economics*, 18, 34-39.

[16] Farrell, J. (1988), 'Communication, Coordination and Nash Equilibrium', *Economics Letters,* 27, 209-214.

[17] Farrell, J., Rabin M. (1996), 'Cheap Talk', *Journal of Economic Perspectives*, 10, 103-118.

[18] Fehr, E., Schmidt K. (1999), 'A Theory of Fairness, Competition and Cooperation', *Quarterly Journal of Economics*, 114, 817-868.

[19] Fehr, E., Tyran, J.-R. (2005), 'Individual Irrationality and Aggregate Outcomes', *Journal of Economic Perspectives,* 19, 43-66.

[20] Frank R.H. (1988), *Passions within Reason: The Strategic Role of Emotions*, New York: Norton.

[21] Geanokoplos, J., Pearce, D., Stachetti, E. (1989), 'Psychological games and sequential rationality', *Games and Economic Behavior*, 1, 60-79.

[22] Hamilton, V.L., Rytina, S. (1980), 'Social Consensus on Norms of Justice: Should Punishment Fit the Crime?' *The American Journal of Sociology*, 85, 1117-1144.

[23] Hoffman, M.L. (1982), 'Development of Prosocial Motivation: Empathy and Guilt', in (N., Eisenberg, ed.), *The development of prosocial behavior*, San Diego, CA: Academic Press.

[24] von Hirsch, A. (2007), 'The Desert Model for Sentencing: Its Influence, Prospects, and Alternatives', *Social Research*, 74, 413-434.

[25] Isaac, M., McCue, K., Plott C. (1985), 'Public Goods Provision in an Experimental Environment', *Journal of Public Economics*, 26, 51-74.

[26] Isaac, M., Walker J. (1988), 'Communication and Free-riding Behavior: the Voluntary Contribution Mechanism', *Economic Inquiry*, 26, 586-608.

[27] Kandel, E., Lazear, E. (1992), 'Peer Pressure and Partnerships', *Journal of Political Economy,* 100, 801-817.

[28] Kaplow, L., Shavell, S. (2007), 'Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System', *Journal of Political Economy,* 115, 494-514.

[29] Kerschbamer, R., Rauchdobler, J., Sausgruber, R., Tyran, J.-R. (2008), 'How Non-binding Agreements Resolve Social Dilemmas' University of Innsbruck, Manuscript.

[30] Lopez-Perez, R. (2008), 'Aversion to Norm-Breaking: A Model', *Games and Economic Behavior*, forthcoming.

[31] Miettinen, T. (2006), 'Pre-play Negotiations, Learning and Nash-Equilibrium', PhD dissertation, University College London.

[32] Milgrom, P., Roberts, J. (1990), 'Rationalizability, Learning and Equilibrium in Games with Strategic Complementarities', *Econometrica,* 58, 1255-1277.

[33] Millar, K.U., Tesser A. (1988), 'Deceptive Behavior in Social Relationships: a Consequence of Violated Expectations', *Journal of Psychology,* 122, 263-273.

[34] Polinsky, A.M.; Shavell, S. (2000), 'The Economic Theory of Public Enforcement of Law', *Journal of Economic Literature* 38, 45-76.

[35] Rabin, M. (1994), 'A Model of Pre-game Communication', *Journal of Economic Theory,* 63, 370-391.

[36] Rotemberg, J. (1994), 'Human Relations in the Workplace', *Journal of Political Economy,* 102, 684-717.

[37] Sally, D. (1995), 'Can I say "bobobo" and mean "There's no such thing as cheap talk"?', *Journal of Economic Behavior and Organization,* 57, 245-266.

[38] Suetens, S. (2005), 'Cooperative and Noncooperative R&D in Experimental Duopoly Markets', *International Journal of Industrial Organization*, 23, 63-82.

[39] Sutter, M. (2008), 'Deception: The Role of Consequences - A Comment', *Economic Journal,* forthcoming.

[40] Tonry, M. H. (ed.) (2001), *Sentencing and Sanctions in Western Countries*, Oxford, UK: The Oxford University Press.

[41] Vanberg, C. (2008), 'Why Do People Keep Promises? An Experimental Test fo Two Explanations', *Econometrica*, forthcoming.