Contracts, Promises, and Norms - An Approach to Pre-play Agreements*

Topi Miettinen†

January 2008

Abstract

We study enforcement of contracts in partnerships. In line with the widely applied and internalized principle of just deserts, we assume that a contract offender who harms the other more suffers a more severe punishment. 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, the principle of just deserts leads to a conflict between Pareto efficiency of the contract and the incentives to abide by it – the more efficient the contract, the weaker the punishment and the incentives to abide by. The opposite is true when actions are strategic complements under specified conditions. With sufficiently strong strategic complements, if contracts can improve the status quo, then a first-best agreement will be abided by. The results have implications for the literature on legal enforcement, and for those on pre-play negotiations in single-shot and repeated games.

JEL Classification C72, C78, K12, Z13

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

---

*This paper is based on chapter 2 of my Ph.D. thesis at University College London. I am grateful to the Yrjö Jahnsson Foundation for financial support. Thanks to Steffen Huck, Philippe Jehiel, Martin Dufwenberg, David Myatt, Birendra K. Rai, Joel Sobel, Christoph Vanberg, and Anthony Ziegelmeyer for comments.

†*Affiliation:* Max Planck Institute of Economics, Germany. *Address:* Kahlaische Strasse 10, D-07745 Jena, Germany. *E-mail:* miettinen@econ.mpg.de.

1 Introduction

In societies across the globe, and throughout centuries, there has been a wide 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) confirm this consensus in the United States in a sociological study. Not only is there societal agreement on this principle of just deserts, but also formal legal codes, since the code of Hammurabi, largely abide by it.

Punishments play a crucial role in the enforcement of obligations in partnerships. In a partnership[1], 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.

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 how violators of those standards will be punished. Both legal and social punishments, in turn, are shaped by justice principles such as that of harm-fitting of 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. 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.[2] As contract efficiency is improved, the harm inflicted by a marginal contract violation increases, and thus a marginal punishment, 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.

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

[1] Radner (1986) defines partnership as one involving an exogenous random component. For simplicity, we consider deterministic partnerships only.

[2] Bulow et al. (1985) introduce and define the concepts of strategic complements and strategic substitutes. Actions are strategic complements, for instance, if the incentive to increase one's action increases in the action of the other.

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. Yet, this recommendation is in sharp contrast with our basic intuitions of justice and thus it poses a challenge to the design of contracts and their enforcement.

Our results bear implications to three strands of literature. First, Becker (1968) and the subsequent literature[3] 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.[4] The current paper extends this literature by allowing deviations from pre-play messages or agreements to be costly.[5] 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. 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.[6] The view that people dis-

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

[4]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.

[5]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.

[6]Models of social norms assume that people have a preference for abiding by norms, of which promise

like breaching even informal mutual agreements is supported by recent experimental evidence which points out that people more strongly prefer stating the truth if lying harms the other more (Gneezy, 2005).

Finally, 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.[7] Nevertheless, if the severity of the punishment is an increasing function of the harm inflicted (and not a function of the agreement per se), 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 paper is organized as follows. Section 2 presents the model. Section 3 has the main results. Sections 4 and 5 discuss the informal pre-play agreement interpretations in one-shot and repeated game settings, respectively. Section 6 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 interpretations of the contract as an informal pre-play agreement in one-shot and repeated games are discussed in sections 4 and 5, respectively. 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.[8] A

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.

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

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

combination of actions is an *action profile* $s = (s_1, s_2) \in S = S_1 \times S_2$. 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. 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.[9]

The lowest Nash payoff of player i is 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 such 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.[10]

Now suppose that 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 *gain from breaching* to i is $g_i : M \times S_i \longrightarrow \mathbb{R}$, $g_i(m, s_i) = u_i(s_i, m_j) - u_i(m)$.

2.2 Punishment

Legal and contractual governance is modeled as follows. If partners have agreed on m, a player will be punished if she deviates unilaterally and her deviation is detected. The magnitude of punishment depends on the harm inflicted to the other and is given by the function $f(h) : \mathbb{R} \to \mathbb{R}^+$. The goal of the paper is to investigate the implications of the class of punishment functions satisfying the following properties.

- $f(h) = 0$, if $h \leq 0$, and $f(\tilde{h}) \geq f(\hat{h})$ if $\tilde{h} > \hat{h} > 0$.
- 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 $h = \lambda \tilde{h} + (1 - \lambda)\hat{h}$.

[9] 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).

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

There is no punishment if no harm is inflicted or the deviation benefits the other. The punishment is strictly positive if the harm inflicted is strictly positive, and weakly increasing for strictly positive levels of harm. Finally, the punishment is weakly convex in the harm inflicted.[11]

Our assumptions allow for a number of possible punishment functions. An example of a punishment 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 crime-fitting applied in the Hammurabi code, incorporating the principles such as "eye for an eye and tooth for a tooth". A fixed punishment

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

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

2.3 The transformed game

Suppose the players have agreed to play m, the punishment function is $f(\cdot)$, and the exogenously given probability of detecting a deviation from an agreement is $\theta \in [0, 1]$. Let $\Gamma(m; \theta)$ denote the non-cooperative interaction following the agreement m. The net payoff to each player depends on the agreement (m), the vector of actions actually played (s), and the probability of detection, (θ). 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

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

Note that there is no punishment if both parties deviate from the agreement.[12]

[11] 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 punishment function can not (see section 6).

[12] This assumption simplifies the model. It is not central for the results. It is also a rather natural assumption since often it is not illegal for a victim of a crime to defend herself. When both deviate, the player's agreed cannot be the right reference, because victim is distorting the reference herself. The assumption ensures that harm need not be measured in the case of mutual deviations.

Given an agreement m, if player i deviates to s_i but player j complies to the agreement, then, using the definition of benefit and harm from breaching, the net payoff to player i can be rewritten as

$$U_i(m_i, m_j, s_i, m_j) = u_i(m) + g_i(m, s_i) - \theta f(h_j(m, s_i)) \tag{4}$$

where the first two arguments of $U_i(\cdot)$ are the agreed actions and the last two entries are the played actions of i and j respectively.

Analogously, the payoff to player j who complies to the agreement can be written as

$$U_j(m_j, m_i, m_j, s_i) = u_j(m) - h_j(m, s_i). \tag{5}$$

2.4 Individually rational and incentive compatible agreements

Let us now define the agreements that we consider the players being able to agree upon. These *acceptable* agreements satisfy two conditions for each player, (i) individual rationality and (ii) incentive compatibility.

An agreement m will be referred to as an *individually rational agreement* for player i if the player prefers the agreement being played to her worst case scenario without an agreement, which in turn provides her lowest Nash payoff. Formally, m is an individually rational agreement if

$$u_i(m) \geq u_i^*, \text{ for } i = 1, 2. \tag{IR_i}$$

The worst Nash equilibrium is naturally the least demanding requirement for individual rationality. Other outside options with higher payoff-criteria could be used equally well.

Although a player would prefer playing an individually rational agreement to playing an underlying game Nash equilibrium, she may not have an incentive to abide by it, if the other abides by it. Let us define player i's *incentive to breach* an agreement m as the difference between the gain from breaching and the punishment, given that the other player does not deviate. Once entered into, an agreement will be *abided* by, if for each of each actions, the player's incentive to breach is non-positive, assuming that the other player does not deviate. Perhaps more simply, the agreed action of i must be a best reply to the agreed action of j. An agreement is *incentive compatible* if

$$B_i(m, s_i; \theta) \equiv g_i(m, s_i) - \theta f(h_j(m, s_i)) \leq 0, \text{ for all } s_i \in S_i, \text{ for } i = 1, 2. \tag{IC_i}$$

The *agreements acceptable to* i can now be defined as $A_i(\Gamma, \theta) = \{m|\ m$ satisfies (IC_i) and $(IR_i)\}$. The *set of acceptable agreements* is defined as the intersection of what is acceptable to each separately, i.e., $A(\Gamma, \theta) = \cap_{i=1,2} A_i(\Gamma, \theta)$. Thus, both parties will not only be willing to voluntarily enter an acceptable agreement, but will have an incentive to abide by it as well.

Notice that, 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 acceptable for any θ in [0,1]. If $\theta = 0$, then only Nash equilibria are acceptable.*

2.5 Preliminaries

Let us first illustrate the above defined concepts in a simple partnership with a prisoner's dilemma structure. By means of this example, let us contrast strategic legal enforcement with the non-strategic one.

Each player decides whether to contribute to the partnership or not. It is efficient that both contribute but it is a strictly dominant strategy not to contribute. A prisoner's dilemma is given in the game matrix in below, where $h_i > u_i > 0$ and $g_i > 0$ for $i = 1, 2$.

	C	N
C	u_1, u_2	$u_1 - h_1, u_2 + g_2$
N	$u_1 + g_1, u_2 - h_2$	$0, 0$

Fig.1. Prisoner's dilemma.

Let us suppose that the punishment takes the simple form of (1) with the Hammurabi rule, $\varphi = 1$. With this specification, player i respects an agreement to contribute, $m = (C, C)$, if and only if the probability of detection is greater than the ratio of the benefit to the harm, $\theta \geq g_i/h_j$. This is player i's incentive compatibility condition.

Both choosing N constitutes the only Nash equilibrium of the underlying dilemma and results in zero payoffs. Thus, both (N,N) (N,C) and (C,C) are individually rational for the row player, since these outcomes result in a non-negative payoff for her. Therefore (C,C) is acceptable if $\theta \geq g_i/h_j$ holds for both players. Otherwise only the Nash equilibrium of the underlying game is acceptable.

Notice that if the benefit is greater than the harm, then the Hammurabi code is unable to deter the violation of a cooperative agreement even if crimes are detected with certainty. Since the benefit of the violation is greater than the harm, a *surplus-maximizing* planner would not strive to prevent the violation in a *non-strategic* model.[13] Yet, in a *strategic* partnership, the lack of sufficient enforcement prevents contracting altogether since an agent, who knows that the other will breach, will either breach herself or will not enter into an agreement in the first place. Although a cooperative agreement is individually rational and each player would gain u_i if cooperation was enforceable, the lack of incentive compatibility leads to a *Pareto inefficient* outcome. As in a non-strategic environment, fitting punishment to the crime implies that not all offenses are prevented. In a strategic environment, this lack of sufficient enforcement will have averse consequences even for a surplus maximizing planner. In the next section, we will establish conditions on strategic settings identifying when the just deserts principle is particularly problematic for efficiency.

3 Partnerships

Let us focus on *finite symmetric 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, $S_i = \{0, ..., n\}$, and we call n the *maximal action*. The actions are strategic complements (substitutes) if an increase in the opponent's action increases (decreases) the marginal payoff.

We will demonstrate two main results. First, there is a conflict between the incentives to abide by an agreement and the Pareto efficiency of the agreement in games with strategic substitutes where the payoffs are monotone in the opponent's action. The incentives necessarily deteriorate as efficiency is improved. Second, this conflict is absent in symmetric games with sufficiently strong strategic complements and monotone payoffs in the opponent's action. Actually, if initially the marginal gain is weakly greater than the marginal punishment, then a first-best contract will be abided by if any agreement which symmetrically improves efficiency is abided by.

3.1 Set-up and preliminaries

We first make some *further assumptions on the underlying game*.

[13] See Polinsky and Shavell (2000, p. 50), for instance.

A1. For any given action of player i, her payoff is increasing in the action of player j.

A2. Each player's payoff is weakly concave in her own action. Or, for all s,

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

A3. Each player's payoff is weakly concave in the action of the opponent. Or, for all s,

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

For simplicity, δ_i and σ_i are constant. The first assumption 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.

To simplify exposition, we now adopt some new concepts.

- For $s \in S$ and for $k \in \mathbb{Z}$, let us call $s + k = (s_1 + k, s_2 + k)$ a symmetric change of actions by k vis-à-vis s.

Notice that s itself does not have to be a symmetric action profile as long as both actions are increased or decreased by the same amount.

- The marginal gain from breaching to player i is $\gamma_i(m_i, m_j) = g_i(m_i, m_j, m_i - 1)$.
- The marginal harm to player i is $\eta_i(m_i, m_j) = h_i(m_i, m_j, m_j - 1)$.

These are the effects on i's UG-payoff and j's UG-payoff, respectively, due to a unit downward deviation from the agreement m by i. Based on these definitions we can define the following.

- For $k \in \mathbb{Z}$, the effect on the marginal benefit, on the marginal harm, and on the agreed payoff due to a symmetric change of agreed actions by k vis-à-vis m are $\gamma_i(m + k) = \gamma_i(m_i + k, m_j + k)$, $\eta_i(m + k) = \eta_i(m_i + k, m_j + k)$ 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 such symmetric changes.

Notice that by the first assumption, the marginal harm, $\eta_j(m)$, is always positive. Marginal gain from breaching, $\gamma_i(m)$, can be positive or negative since own-action

monotonicity is not assumed. However, player i will not be punished if she makes both better off by deviating. Consequently, each player's agreed payoff must necessarily be non-increasing[14] at any agreement which is acceptable. We denote this feasible set of agreements by M^F

$$M^F = \cap_{i=1,2} M_i^F \text{ where } M_i^F = \{m | \gamma_i(m_i + 1, m_j) \leq 0\}. \tag{6}$$

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 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.[15]

Let us now proceed towards our main results. We wish to study how the incentives to breach are affected if an agreement is altered so as to improve efficiency. We are particularly interested in comparing this impact in games with strategic complements, on the one hand, and games with strategic substitutes, on the other hand.

3.2 Strategic substitutes

Let us now assume that actions are strategic substitutes. By definition, actions are strategic substitutes if for all s, $u_i(s_i, s_j) - u_i(s_i, s_j - 1) - [u_i(s_i - 1, s_j) - u_i(s_i - 1, s_j - 1)]$

[14] Except for $m_i = n$ of course.

[15] Notice that it is crucial that we assume that f is a weakly convex function. Otherwise non-marginal deviations might pay off although a marginal deviation does not.

$= \phi_i \leq 0$. Let us 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.[16] 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 player dislikes increments of her contribution above the equilibrium since, within the feasible set of agreements M_i^F, this necessarily reduces payoff. Moreover, since payoffs are concave, the marginal incentive to breach, $\gamma_i(m)$, increases if her own agreed action is increased. On the other hand, the opponent likes a player's contribution being increased 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 the player's own agreed action unambiguously increases the incentive to breach.

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

Lemma 3

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

With strategic substitutes, $\phi < 0$, the effects in lemma 3 on the marginal incentive to breach 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 4 *Let actions be strategic substitutes. If efficiency is improved vis-à-vis an interior equilibrium due to a symmetric change of actions, then the marginal gain from breaching*

[16] See lemma 13 in the mathematical appendix

increases and the marginal harm decreases. Formally, let s^* be an interior equilibrium. If $u_i(s^* + k) - u_i(s^*) > 0$ for $k \in \mathbb{Z}$ and for $i = 1, 2$, then $\gamma_i(s^* + k) > 0 \geq \gamma_i(s^*)$, and $\eta_i(s^* + k) < \eta_i(s^*)$ for $i = 1, 2$.

3.3 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. In lemma 3, this impact was shown to decrease the marginal incentive to breach within M_i^F. Yet, the effect of increasing the *opponent's* agreed action is now the opposite.

Proposition 5 *Let actions be strategic complements. Then $i's$ marginal incentive to breach is non-decreasing in m_i and non-increasing in m_j in the feasible set.*

The proposition establishes that, if an agreed action is unilaterally changed so as to increase a player's agreed payoff, the player's incentive to breach decreases. Thus, there is no conflict between a given player's agreed payoff and the player's incentive to abide by the agreement when either action is changed unilaterally. In games with strategic complements, the principle of harm-fitting of punishments thus implies another principle which has some flavor of reciprocity: deviators of more generous agreements will be punished more severely.

Yet, only with strategic complements and not with strategic substitutes, does the first reciprocal principle imply the second principle. Moreover, it only holds for unilateral changes in the terms of the agreement - the Pareto efficiency of an agreement could be at variance with the incentives to abide by it if both agreed actions are changed. Indeed, 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 substitutes, 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 more in detail, we can alternatively decompose the effect on incentives into the benefit-effect (lines 1 and 3 in lemma 3) and the harm effect (lines 2

and 4 in lemma 3). When $\phi + \delta \geq 0$, the marginal gain from breaching is decreasing as efficiency is improved. In this case, strategic complementarity is very strong - so strong that maximal actions are efficient and, moreover, they constitute a Nash equilibrium of the underlying game. Therefore, by proposition 1, the maximal actions are acceptable.[17] 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 equilibrium (according to Pareto ranking).

Theorem 6 *Let actions be strategic complements. Let s^* be the most efficient interior underlying game equilibrium. If*

- *the underlying game payoffs satisfy $\phi + \sigma \geq 0$*
- *for each $i = 1, 2$ there are k_i', k_i'' and $\overline{k}$ such that $0 \leq k_i'' < k_i' < \overline{k}$, and player i would breach an agreement $s^* + k_i''$ and would not breach an agreement $s^* + k_i'$ and $s^* + \overline{k}$ is an efficient action profile.*

then $s^* + \overline{k}$ is acceptable.

When strategic complementarity is weaker, the marginal gain from breaching is increasing as efficiency is improved, $\delta + \phi < 0$. If complementarity is so weak that it does not even dominate the concavity effect the increase in the action of the opponent (let us call this *the opponent-concavity effect*), $\sigma + \phi < 0$, then also the marginal harm is decreasing 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, $\phi + \sigma \geq 0$, then the marginal harm is non-decreasing and acts as an opposing force to the marginal benefit as efficiency is improved.[18] 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

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

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

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

Notice that we need some rather unrestrictive assumptions to establish our result. Our assumptions guarantee that, if the increasing marginal harm eventually balances out an initially too tempting breaching incentive, then this balance will hold for any agreement which is more efficient than this cut-off agreement.

4 Informal agreements and social pressure

The legal enforcement of formal contracts has a natural analog. The agreement can be interpreted as an informal one. In this case, the punishment can be thought of as social pressure or social punishment. Pressure and punishments are carried out by the offender or outside observers. Alternatively, the punishment could be intrinsic - feelings of guilt and shame by the offender herself.

Our just deserts assumption typically holds in the contexts of informal agreements. First, the principle that "punishment should fit the crime", is a very commonly held social justice principle (Hamilton and Rytina, 1980). Commonly shared social codes regulate how severely violators of norms should be punished. People conform to those rules of punishment for the very same intrinsic and extrinsic reasons as they choose not to transgress social norms in the first place. As far as the intrinsic motivation is concerned, Hoffman (1982), for instance, suggests that guilt has its roots in a distress response to the suffering of others which reflects internalized social norms. Moreover, Gneezy (2005) finds support for the view that private preferences for breaching promises reflect the just deserts principle: his experiment suggests that people trade off the benefits of lying against the harm that lying inflicts on the opponent.[19]

Let us pursue the intrinsically motivated approach. For concreteness and to reflect the emotion-based intrinsic motivation, let us call the punishment for defecting as *guilt* in the rest of this section. Those preferring a mindless approach can carry along the analog interpretation as mere preference or as a punishment by victim or third parties. In the intrinsic context, the parameter θ is an individual-specific one measuring sensitivity to pressure or proneness guilt, not merely a probability of detection. Particularly, in the case of guilt, probability of detection plays no role since the punishment is entirely intrinsic and a transgressor surely knows that she has transgressed the agreement. Whichever the source of the breaching cost, there is a reason to believe that there

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

is an abundance of contexts where informal promises about intentions are not merely cheap talk (Sally, 1995) but rather breaking a promise induces a cost.

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 (1987), 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 common knowledge. As opposed to Farrell, the model in this paper does not explicitly model the communication. Yet, nothing prevents one doing that - any negotiation protocol could be added to analyze its effect on agreements and play.[20]

The results of the previous sections fully carry over to the alternative interpretation of agreements by informal pre-play negotiations. Moreover, in the appendix, the informal agreements model is extended to allow guilt to respond reciprocally to generosity. This means that guilt is allowed to depend not only on the harm inflicted on the opponent but also it is allowed to increase in the player's agreed payoff. The more the opponent agrees to give, the more guilty a player may feel about letting her down. This approach allows us to prove a result analogous to theorem 6. We can also derive a corollary which strengthens the result: if anything more efficient than a unique UG equilibrium is acceptable, then an efficient agreement is also acceptable. The approach also allows us to strengthen proposition 2. Whereas before, marginal incentive compatibility was equivalent to incentive compatibility, in the present context marginal incentive compatibility is equivalent to acceptability.

Some assumptions that have been made in Sections 2.2 and 2.3 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. As pointed out in Section 2.5, this guarantees that marginal violations of the agreement will have non-marginal implications. Bicchieri (2006, chapter 1) discusses in length why conditional conformism captured by this feature is necessary in any reasonable account of social norms.[21] The simple idea hidden in conditional conformism when applied to partnerships is that people do not like to be exploited suckers.

[20] Miettinen (2008) does this in the setting of the alternating offer protocol.

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

5 Repeated games

We have studied one shot games and implied punishments. Yet, the conclusions of the model will continue to hold if the interaction is repeated and punishments are endogenous to the interaction. The principles of just deserts may be assumed to restrict the allowable punishment paths, and even renegotiations, in a more descriptive, rather than rationally prescriptive, account of repeated games analyzing applied social situations and relational contracts.

The folk argument of infinitely repeated games illustrates that typically a large number of enforceable agreements are available for partners engaging in a repeated interaction. 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[22], can be sustained in a Nash equilibrium of an infinitely repeated game. Punishing a deviator by holding her to her reservation payoff forever, if she deviates from m, suffices to provide an incentive to abide by the agreement.

Abreu (1988) illustrates that in a perfect equilibrium the most severe punishments have a stick-and-carrot structure. The punishment path begins with a harsh punishment phase and then proceeds towards a reward for punishing. Thus holding the other to her reservation payoff for infinitely long payoff is not feasible. The stick and carrot punishment is, nevertheless, stronger than punishing a player with her lowest stage game Nash equilibrium payoff of that player, u_i^*, (Friedman, 1971).[23]

Whichever equilibrium concept one uses, the just deserts principle requires punishing less harmful deviations with a less-than-maximal punishment. Thus enforcement of low-harm deviations is weaker than maximal under just deserts[24]. The set of enforceable agreements is a subset of that identified a given folk theorem. Our main results hold even in the repeated game settings if the severity of the punishment is restricted to be an increasing function of the harm on the other. Again, if the stage game has strategic substitutes, the efficiency of the pre-play agreement stands in contradiction with the incentives to abide by it and, yet, such a discord is circumvented if actions are sufficiently strong strategic complements.

---

[22] This is also called player i's minmax payoff, formally $\underline{u_i}=\min_{s_j}\{\max_{s_i} u_i(s_i,s_j)\}$.

[23] Thus, in the repeated game setting, the individually rational payoff may be set even below u_i^*.

[24] Binmore (1998) provides an inspiring account of social norms as an equilibriun selection device in infinitely repeated games. When the social contract applies the just deserts principle for the punishment paths, there is less equilibria that social norms need to select from.

6 Discussion

The model in this paper illustrates the implications of "punishments that fit the harm" on the enforcement of bilateral partnerships. We show that, 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 neither would abide by and another more efficient agreement that both would abide by, then both are willing to abide by a *first best* agreement.

Our approach is in line with results from public good experiments where communication significantly increases contribution levels.[25] Isaac and Walker (1988) adopt a constant-returns-to-scale technology in a voluntary public good provision game implying a setting with weak strategic complements. They find a strong positive effect of communication on efficiency. Average contribution levels are practically first-best efficient. In a design with decreasing returns to scale implying strict strategic substitutes, they find that communication increases contributions much less. This latter result is backed up with a similar finding by Isaac et al. (1985) in a design with decreasing returns to scale.[26]

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. The incentives to abide by agreements close to an interior group optimum are particularly weak since actions are strategic substitutes. On the other hand, the incentives to respect are the strongest close to a boundary group optimum since actions are strategic complements. Notice that, in order to account for this difference, it is crucial that the guilt is increasing and weakly convex in the harm on the other. A constant cost of deviating[27] (Ellingsen and Johanneson, 2004), for instance, implies that cost is concave in the harm on the other overall, due to the discontinuity at zero. Therefore a constant cost cannot account for the differences in the interior and

[25] Alternative accounts explaining stronger cooperation under strategic complements than substitutes hinge upon bounded rationality (Haltiwanger and Waldman, 1985; Akerlof and Yellen, 1985; Fehr and Tyran, 2005) or evolution (Bester and Güth, 1999). Agents in our model are rational although they may have other-regarding motivations. Moreover, our approach accounts for why negotiations and agreements may particularly reinforce cooperation.

[26] These two studies unlike many others allow subjects to play repeatedly and learn about the game.

[27] See equation (2).

boundary group optimum experiments.

There is further experimental support for our theoretical prediction that informal agreements and norms tend to enforce efficient play better when actions are strategic complements than substitutes. Suetens (2005) finds that communication induces cooperation in an R&D environment which is characterized by strategic complements but that there is no cooperation in an environment with strategic substitutes. Suetens and Potters (2007) review data from four duopoly experiments and find that, with substitute products, there is more tacit collusion in Bertrand duopolies (strategic complements) than in Cournot duopolies (strategic substitutes). Potters and Suetens (2006) design a controlled experiment to compare cooperation in designs with strategic complements and substitutes and find that there is more tacit collusion with strategic complements than with strategic substitutes.

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. Ellingsen and Johanneson (2004), Bicchieri (2006), Lopez-Perez (2008), and Dufwenberg and Battigalli (2007) propose alternatives to the model of guilt in this paper[28].

The former three are more traditional outcome-based models where the agreement is considered as part of the outcome. The present model extends the existing outcome-based models of guilt by allowing guilt to be increasing in the inflicted harm. This feature is crucial for our results.

The last of these papers models 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[29]. In that interpretation the agreement enters the utility function 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.

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

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

In sections 4 and 5, we studied informal pre-play agreements in one-shot and repeated games, respectively. The pre-play negotiation protocol was unmodelled. 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 or convention - 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.

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.

7 Appendix A, the role of generosity

In this appendix, we extend the model by introducing an additional property to the punishment. The primary interpretation of the agreement is the informal one and the punishment should be considered as the intrinsic disutility of feeling guilty about transgressing the informal agreement. We assume that guilt is increasing in the generosity of the agreement. If the opponent keeps her promise and moreover the payoff is high if I keep my word, too, then the opponent is not only sticking to the promise but she is also generous. We assume that guilt about breaching the agreement and not reciprocating will induce stronger guilt than if the agreement had been less generous. We will prove a result analogous to theorem 6 and a corollary which strengthens the result: if anything more efficient than a unique UG equilibrium is acceptable, then an efficient agreement is also acceptable. This approach also allows us to strengthen proposition 2. Whereas before, marginal incentive compatibility was equivalent to incentive compatibility, in the present context marginal incentive compatibility is equivalent to acceptability.

To introduce the generosity-effect formally, we can adopt a natural measure for the generosity of the agreement: The *agreed payoff* indicates how much more than u_i^* the player gets[30] 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. The parameter θ_i would reflect the player's sensitivity to social pressure or her proneness to feelings of guilt. This proneness is naturally individual specific and can take any non-negative value. Notice yet that even partners who were infinitely prone to guilt could not agree on everything since each party would have no trouble making efficiency-improving unilateral deviations.

Lemma 7 *For any m, if there is a player i such that there exists s_i such that $u_i(s_i, m_j) > u_i(m)$ and $u_j(s_i, m_j) \geq u_j(m)$ then m is not acceptable for any θ.*

Proof. By assumption, f(h$_j(m, s_i)) = 0$ if $h_j(m, s_i) < 0$. By assumption, there is j such that, $h_j(m, s_i)$=$u_j(m) - u_j(m_j, s_i) < 0$ and $g_i(m, s_i) = u_i(s_i, m_j) - u_i(m) > 0$. Thus for any θ_i, $B_i(m, s_i, \theta_i) = g_i(m, s_i) - \theta_i f(h_j(m, s_i)) = u_i(s_i, m_j) - u_i(m) - \theta_i f(0) > 0$. Therefore (IC_i) is violated and $m \notin A_i(\Gamma, \theta_i)$ and thus $m \notin A(\Gamma, \theta)$. ■

[30] 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.

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. Moreover, 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 we assume that a partner has no trouble behaving opportunistically.

$$\begin{array}{ll} f(v_i, h_j) > 0, & \text{if } h_j > 0, v_i > 0 \\ f(v_i, h_j) = 0, & \text{if } h_j \leq 0 \text{ or } v_i \leq 0 \end{array}. \tag{7}$$

An example of a suitable guilt function is

$$f(v_i(m), h_j(m, s_i)) = \max\{v_i(m), 0\}^{\gamma} \max\{h_j(m, s_i), 0\}^{\varphi}. \tag{8}$$

The entire game preferences of this form with $\gamma = \varphi = 1$ are closely related to tractable preferences of Cox-Friedman-Gjerstad (2006). 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 Cox-Friedman-Gjerstad preferences. 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-Friedman-Gjerstad (2006). Thus, our model can be broadly considered as a tractable model of guilt.

When accounting for the additional properties of guilt, we can show that a version of theorem 6 holds.

Theorem 8 *Let s^* be the most efficient interior UG equilibrium. If*

- *$u_i(s + k)$ is convex in k.*
- *f is weakly convex in v_i and supermodular*[31] *in its arguments*
- *for each $i = 1, 2$ there are k_i' and k_i'' such that $0 \leq k_i'' < k_i'$ and player i would breach an agreement $s^* + k_i''$ and not breach an agreement $s^* + k_i'$, that is $\beta_i(s^* + k_i'', \theta_i) \geq 0$ and $\beta_i(s^* + k_i', \theta_i) \leq 0$,*

then a Pareto-efficient agreement is acceptable.

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

In this framework of informal agreements, we can prove even a more powerful result stated in corollary 9. If there exists an acceptable informal agreement that improves efficiency vis-à-vis a status quo with a unique equilibrium, then a Pareto efficient agreement is acceptable.

Corollary 9 *Let s^* be the unique underlying game equilibrium. Let $\gamma_i(s^*) = 0$ for $i = 1, 2$. If*

- *$u_i(s + k)$ is convex in k*
- *f is weakly convex in v_i and supermodular in its arguments*
- *for each $i = 1, 2$ there is k_i' such that the marginal incentive to breach is non-positive at $s^* + k_i'$ $\beta_i(s^* + k_i', \theta_i) \leq 0$,*

then a Pareto-efficient agreement is acceptable.

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.

Notice also that in this setup, we can almost *characterize* acceptability in terms of marginal incentive compatibility.

Proposition 10 *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. $m_i \notin \{0, n\}$. An action profile is acceptable for i if and only if i's marginal incentive to breach is non-positive.*

Proof. The result follows directly from lemma 11 and lemma 12 below and the fact that $A(\Gamma, \theta) = \cap_{i=1,2} A_i(\Gamma, \theta_i)$. ■

This is a stronger result than proposition 2 which shows that marginal incentive compatibility is equivalent to incentive compatibility, not acceptability. The result is mainly driven by the fact that when a player's agreed action is not an underlying game best-reply to the agreed action of the opponent, then incentive compatibility implies individual rationality.

Lemma 11 *Let m_i be an underlying game best-reply to m_j. Then m is acceptable for i given θ_i if and only if (IR_i) holds.*

Let m_i not be an underlying game best-reply to m_j. Then m is acceptable for i given θ_i if and only if (IC_i) holds.

Proof. Let m_i be an underlying game best-reply to m_j. Or formally, for all s_i, $u_i(m_i, m_j) \geq u_i(s_i, m_j)$. Since enforcement cost can only be negative, it is then also true that for all s_i, $u_i(m_i, m_j) \geq u_i(s_i, m_j) - f(v_i(m), h_j(m, s_j))$. Thus, by the definition of the incentive to breach, for all s_i, $B_i(m, s_i; \theta_i) \leq 0$ which is the definition of incentive compatibility. Thus, $m \in A_i(\Gamma, \theta_i)$ if and only if m is individually rational.

Now consider the second claim. If $m_i \notin BR_i(m_j)$ then there is s_i' such that $u_i(s_i', m_j) > u_i(m_i, m_j)$. Suppose now that the agreement is incentive compatible. Then by definition, for all s_i, $B_i(m, s_i; \theta_i) \leq 0$ implying that also $B_i(m, s_i'; \theta_i) \leq 0$. Thus, $f(v_i(m), h_j(m, s_i'))$ $\geq u_i(s_i', m_j) - u_i(m_i, m_j)$. But since the right-hand side is positive by assumption, also $f(v_i(m), h_j(m, s_i')) > 0$. Yet, by assumption this can only hold if $v_i(m) > 0$. Thus individual rationality holds and $m \in A_i(\Gamma, \theta_i)$. Suppose now that m is acceptable. Then by definition incentive compatibility holds. ■

8 Appendix B, proofs

Here we give proofs for the extended model where cost is a function of generosity of j, v_i, in addition to the harm (appendix A). 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.

8.1 Proof of proposition 2

Lemma 12 returns proposition 2 in the text (where f is a function of h_j only) as a special case.

Lemma 12 *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 $\beta_i(m, \theta_i) \leq 0$.*

Proof. 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 {1}, the 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 the 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. ■

8.2 Proof of lemma 3

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

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

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

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

8.3 Proof of proposition 5

The marginal incentive to breach reads

$$\beta_i(m_i,m_j)=\gamma_i(m_i,m_j)-\theta_i f(v_i(m_i,m_j),\eta_j(m_i,m_j)).$$

In this expression, $\gamma_i(m)$ is non-decreasing in m_i and $\eta_j(m)$ is non-increasing in m_i by lemma 3. Also, $u_i(m_i+1,m_j)-u_i(m_i,m_j)\leq 0$ since m is acceptable for i. This implies

that $u_i(s)$ is non-increasing in s_i at m. But f is non-decreasing in both arguments. Thus, $\beta_i(m_i, m_j)$ is non-decreasing in m_i.

On the other hand, $\gamma_i(m_i, m_j)$ is non-increasing in m_j and $\eta_j(m_i, m_j)$ is non-decreasing in m_j by lemma 3. Also, u_i is increasing in m_j by assumption. But f is non-decreasing in both arguments. Thus, $\beta_i(m_i, m_j)$ is non-increasing in m_i. ■

8.4 Proof of theorem 4

Lemma 13 *Let actions be strategic substitutes. If* $u_i(s_i^* + k, s_j^* + k) - u_i(s^*) > 0$ *for* $i = 1, 2$, *then* $k > 0$

Proof. Since the equilibrium is interior, by the concavity of the payoffs in own actions, we must have $\gamma_i(s^*) \leq 0$ and $\gamma_i(s_i^* + 1, s_j^*) \geq 0$ for $i = 1, 2$. Since {3'} holds, by lemma 3, $\gamma_i(s_i^*, s_j^* - 1) < 0$ and therefore, by {1}, $u_i(s^*) - u_i(s^* - 1) = u_i(s^*) - u_i(s_i^*, s_j^* - 1) - \gamma_i(s_i^*, s_j^* - 1) > 0$. More generally, by lemma 3 and assumptions {3'} and {1}, $\gamma_i(s_i^* + 1 - k, s_j^* - k) < 0$ for $k \geq 1$ and thus, for each such $s^* - k$, $u_i(s^* - k + 1) - u_i(s^* - k) > 0$ for $i = 1, 2$. Therefore, if $u_i(s_i^* + k, s_j^* + k) - u_i(s^*) > 0$ for $i = 1, 2$, then $k > 0$. ■

Proof of the theorem

Since for all s, $\phi_i < 0$ and $\delta_i, \sigma_i \leq 0$, by lemma 3, $\gamma_i(s^* + k) > \gamma_i(s^*)$ and $\eta(s^* + k) < \eta(s^*)$ for $i = 1, 2$. if $k > 0$. The latter holds by lemma 13. ■

8.5 Proof of theorem 8

Lemma 14 *Let* $0 > \phi + \delta$. *Let* s^* *be the most efficient interior equilibrium in the underlying game. Then*

- *the marginal gain from breaching is negative at* s^*, *i.e.* $\gamma_i(s^*) \leq 0$
- *the marginal gain from breaching is positive at* $(s_i^* + 1, s_j^*)$, *i.e.* $\gamma_i(s_i^* + 1, s_j^*) \geq 0$
- s^* *is Pareto-preferred to any* s *such that* $s_i \leq s_i^*$ *for* $i = 1, 2$

  Moreover, if $s^* + 1 \neq (n, n)$ *then*

- *the marginal gain from breaching is positive at* $(s_i^* + 1, s_j^* + 1)$, *i.e.* $\gamma_i(s_i^* + 1, s_j^* + 1) > 0$

Proof. The first two properties are satisfied since s^* is an interior equilibrium and the payoff is concave in player i's own action. The third property follows since the payoff is increasing in the opponent's action and s_i^* is a best-reply to s_j^* and thus for

any s and for $i = 1, 2$, $u_i(s^*) \geq u_i(s_i, s_j^*) > u_i(s_i, s_j)$. For the fourth property, suppose that $\gamma_i(s_i^* + 1, s_j^* + 1) \leq 0$. Now $\gamma_i(s_i^* + 1, s_j^*) \geq 0$ implies, by lemma 3, that $\gamma_i(s_i^* + 2, s_j^* + 1) > 0$ since $-\delta > \phi$. By symmetry this holds also for j. Thus $s^* + 1$ is an equilibrium and, by the assumption $s^* + 1 \neq (n, n)$, it is an interior equilibrium. Moreover, since u_i is increasing in s_j, by theorem 7 in Milgrom, Roberts (1990), $u_i(s^* + 1) > u_i(s^*)$. (Milgrom and Roberts show that, when payoff is increasing in the opponent's action and actions are strategic complements, the equilibria can be Pareto-ranked so that the equilibrium with the highest contributions is the Pareto-best.) This is a contradiction to the assumption that s^* is the most efficient interior equilibrium. ■

Lemma 15 *If $s^{PB} = (s + \overline{k})$ where $\overline{k} = \arg max_{k \in Z} u(s + k)$ and $s_i = s_j$, that is s^{PB} is Pareto-best among symmetric strategies, then there is no (s_i', s_j') such that $u_i(s') > u_i(s^e)$ for $i = 1, 2$.*

Proof. Let without loss of generality, $s_j' < s_i'$ and $s_i' - s_i^{PB} = k$. Then $u_i(s^{PB}) > u_i(s^{PB} + k) = u_i(s_i', s_i') > u_i(s_i', s_j')$ since the payoff is increasing in the action of the opponent. Thus s^{PB} is efficient. ■

Lemma 16 *Let s^* be a symmetric interior underlying game equilibrium. If $s^* + \overline{k}$ is efficient, then $\overline{k} > 0$. There exists $\overline{k}$ such that $s^* + \overline{k}$ is efficient.*

Proof. Let the underlying game second differences satisfy $\phi + \delta \geq 0$. By lemma 14, the marginal incentive to breach satisfies $\gamma_i(s^*) \leq 0$ and $\gamma_i(s_i^* + 1, s_j^*) \geq 0$. Now since $\phi + \delta \geq 0$, by lemma 3, $\gamma_i(n) \leq 0$ and thus (n, n) is an (boundary) equilibrium. Moreover (n, n) is the Pareto-optimal profile since the payoff is increasing in the opponent's action and n is a best-reply to n and thus for any s for $i = 1, 2$, $u_i(n, n) \geq u_i(s_i, n) > u_i(s_i, s_j)$.

Suppose now that $\phi + \delta < 0$. Let us argue that the profile that maximizes each underlying game payoff along the diagonal, $\max_{k \in Z} u(s^* + k)$ where $s_i^* = s_j^*$, is $s^* + \overline{k}$ where $\overline{k} > 0$. First let the second differences satisfy $\sigma + \delta + 2\phi \geq 0$, that is $u_i(s + k)$ is convex in k. By the third property in lemma 14, $u_i(s^*) - u_i(s^* - k) > 0$ for any $k > 0$. Therefore (n, n) maximizes each underlying game payoff along the diagonal. Let $\phi + \delta < 0$ still hold and suppose alternatively that $\sigma + \delta + 2\phi < 0$. Then $u_i(s + k)$ is strictly concave in k. By lemma 14, $u_i(s^*) > u_i(s^* - 1)$ for $i = 1, 2$. Since the strategy set is finite, a maximizer $s^* + \overline{k}$ along the diagonal exists and it satisfies $\overline{k} > 0$. Finally whether we have $\sigma + \delta + 2\phi \geq 0$ or $\sigma + \delta + 2\phi < 0$, by lemma 15, the profile that maximizes the payoff along the diagonal is efficient. ■

Lemma 17 *Let the underlying game payoff second differences satisfy $\phi+\delta<0$ and $\phi+\sigma\geq 0$. Let s be symmetric, that is $s_i=s_j$. 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 $k>0$.*

Proof. By lemma 3, 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 $\eta\geq 0$.

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,

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

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 acceptable. ■

Proof of the theorem 6

Let the underlying game payoff second differences satisfy $\phi+\delta\geq 0$. By lemma 14, $\gamma_i(s^*)\leq 0$ and $\gamma_i(s_i^*+1,s_j^*)\geq 0$. Now since $\phi+\delta\geq 0$, by lemma 3, the marginal incentive to breach is non-positive at the maximal action, $\gamma_i(n)\leq 0$ and thus (n,n) is an equilibrium. Moreover (n,n) is the Pareto-optimal profile since the payoff is increasing in the opponent's action and n is a best-reply to n and thus for any s for $i=1,2$, $u_i(n,n)\geq u_i(s_i,n)>u_i(s_i,s_j)$. Finally by proposition 1, (n,n) is acceptable since (n,n) is an underlying game equilibrium.

Let $\phi+\delta<0$ hold. By assumption, for each $i=1,2$, there are k_i' and k_i'' such that $0\leq k_i''<k_i'$ and such that $\gamma_i(s^*+k_i'')\geq\theta_i f(\eta_j(s^*+k_i''))$ and $\gamma_i(s^*+k_i')\leq\theta_i f(\eta_j(s^*+k_i'))$. If $\phi+\sigma\geq 0$, then the marginal harm, $\eta_j(s^*+k)$, is non-decreasing in k. Moreover, we can apply lemma 17. By assumption $k_i'<\overline{k}$ for $i=1,2$. Thus, $s^*+\overline{k}$ is acceptable. ■

Proof of theorem 8

The proof of theorem 8 is along the lines of the proof above. Yet, instead of applying lemma 17 one should apply 18 below. By assumption $u_i(s^* + k)$ is convex in k (that is, $\sigma + \delta + 2\phi \geq 0$). Thus, the efficient profile $s^* + \overline{k}$ is one where both choose the maximal action, (n, n), and thus necessarily $\overline{k} \geq k_i'$ for $i = 1, 2$.

Proof of the corollary 9

If the underlying game payoff second differences satisfy $\delta + \phi \geq 0$ and $\gamma_i(s^*) = 0,\ (n, n)$ is an equilibrium. Thus, we have a contradiction. If $\delta + \phi < 0$, no symmetric action profile where $s_i \leq s_i^*$ is acceptable. To see this notice that, since $\gamma_i(s^*) = 0$, it must hold that $\gamma_i(s) < 0$ by lemma 3. Therefore, if $\gamma_i(s_i + 1, s_j) \geq 0,\ s$ is an equilibrium contradicting the uniqueness of equilibria. Thus $\gamma_i(s_i+1, s_j) < 0$ implying that $s \notin M^F$. Therefore there is $k_i' > 0$ such that $s^* + k_i'$ is acceptable 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 8. ■

8.6 Proof of lemma 18

Lemma 18 *Let the underlying game payoff second differences satisfy $\phi + \delta < 0$ and $2\phi + \delta + \sigma \geq 0$. Let s be symmetric, that is $s_i = s_j$. Let $u_i(s) - u_i(s - 1) \geq 0$. Let g satisfy {5}. 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 $k > 0$.*

Proof. $\delta + 2\phi + \sigma \geq 0$ and $\phi + \delta < 0$ implies that $\phi + \sigma \geq 0$. Then, by lemma 3, 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 acceptable. ■

References

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

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

[3] Akerlof, G., Yellen, J. (1985), 'Can Small Deviations from Rationality Make Significant Differences to Economic Equilibria?' *American Economic Review,* 75, 708–20.

[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] Bester, H., Güth, W. (1999), 'Is Altruism Evolutionarily Stable?' *Journal of Economic Behavior and Organization*, 34, 193-209.

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

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

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

[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] Haltiwanger, J.C., Waldman, M. (1985), 'Rational Expectations and the Limits of Rationality: An Analysis of Heterogeneity' *American Economic Review,*. 75, 326–40.

[23] 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.

[24] 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.

[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] 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.

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

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

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

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

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

[33] Potters, J, Suetens, S. (2006), 'Cooperation in Experimental Games of Strategic Complements and Substitutes', Center Discussion Paper No. 2006-48. Tilburg University.

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

[35] Radner, R. (1986), 'Repeated Partnership Games with Imperfect Monitoring and No Discounting', *Review of Economic Studies*, 53, 43-57.

[36] 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.

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

[38] Suetens, S., Potters, J. (2007), 'Bertrand Colludes More Than Cournot', *Experimental Economics*, 10, 71-77

[39] Vanberg, C. (2008), 'Why Do People Keep Promises? An Experimental Test fo Two Explanations', Jena Economic Research Paper