Promises and Conventions - An Approach to Pre-play Agreements*

Topi Miettinen†

November 2006

Abstract

Experiments suggest that communication increases the contribution to public goods (Ledyard, 1995). There is also evidence that, when contemplating a lie, people trade off their private benefit from the lie with the harm it inflicts on others (Gneezy, 2005). We develop a model of bilateral pre-play agreements that assumes the latter and implies the former. A preference for not lying provides a partial commitment device that enables informal agreements.

We establish some general properties of the set of possible agreements in normal form games and characterize the smallest and largest such set. In symmetric games, pre-play agreements crucially depend on whether actions are strategic complements or substitutes. With strategic substitutes, commitment power tends to decrease in efficiency whereas the opposite may be true with strategic complements. Also this finding is supported by experimental evidence.

JEL Classification C72, C78, Z13

KEYWORDS: pre-play negotiations, communication, social norms, agreements, guilt

*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. Also, I am very grateful to Steffen Huck and Philippe Jehiel for advice, encouragement, suggestions, and discussions. Finally, I would like to thank Martin Dufwenberg, Daniel Friedman, Antonio Guarino, Klaus Helkama, Ed Hopkins, Klaus Kultti, Mikko Leppämäki, Vittoria Levati, Liisa Myyry, Matthew Rabin, Francesco Squintani, Emma Tominey, Birendra Kumar Rai, Joel Sobel, Pekka Sääskilahti, Hannu Vartiainen, Juuso Välimäki, Georg Weiszäcker and the seminar participants at Copenhagen, Essex, Heidelberg, Helsinki, Max Planck Institute in Jena, St. Andrews, UCL, SAE 2005, Game Theory Festival at Stony Brook, Zeuthen Workshop in Copenhagen, SING2 in Foggia and ESA European Meeting 2006 in Nottingham. All errors are mine.

†*Affiliation:* Max Planck Institute of Economics, Germany. RUESG, Dept of Economics, University of Helsinki and HECER, Finland. *Address:* Kahlaische Strasse 10, D-07745 Jena, Germany. *E-mail:* miettinen@econ.mpg.de.

There is no commonly honest man ...who does not inwardly feel the truth of the great stoical maxim, that for one man to deprive another unjustly to promote his own advantage by the loss or the disadvantage of the another, is more contrary to nature, than death, than poverty, than pain, than all the misfortunes which can affect him, either his body, or his external circumstances.

-Adam Smith (The Theory of Moral Sentiments, p. 159, 2002 (1759))

1 Introduction

Ray and Cal have a magic pot and ten dollars each. Each dollar put into the pot gives $\frac{3}{4}$ dollars to both of them. Ray and Cal have to decide how many dollars to put into the pot and how many to keep to themselves. Ray figures that, whatever Cal puts into the pot, for each dollar he puts into the pot, he gets only $\frac{3}{4}$ dollars back and, hence, should put nothing into the pot.

Before they decide, they can talk to each other. They may agree on how many dollars each of them will put into the pot. The agreement is not binding. Yet, having talked to Cal for a while, he seems like a nice guy to Ray. Ray starts to think that he would feel bad if he lied about how many dollars he will put into the pot. He also figures that Cal may well think similarly about him. Eventually, Ray and Cal agree on putting ten dollars each into the pot and neither violates the agreement.

Most people would think that the story above is vaguely plausible but doubt that such magic pots exist. An economist is certain about the existence of the magic pot, but has doubts whether people care about inflicting harm on the other by not doing as agreed.

Two findings in experimental economics give a reason to believe that the magic pots and the dislike to breach oral agreements are worth taking seriously: First, communication increases contributions in public good games (Ledyard, 1995). Second, if people lie, they tend to dislike it; and they seem to dislike it more if they inflict more harm on others by doing so (Gneezy, 2005). In the public good games, agreeing to contribute more than one actually intends to contribute is a lie which harms others. Therefore, deviating from the promise is less profitable and promises to contribute may be credible. Thus, assuming the latter finding (dislike breaching if harming others) provides an explanation for the former finding (increased contributions in public good games).

This paper presents an approach to pre-play agreements by negotiations, conventions or social norms. We assume that people may feel guilty when lying about their intentions, breaching informal agreements or transgressing social norms. Moreover, we assume that the guilt cost due to a transgression is higher if someone inflicts more harm on the other by doing so. We show that an agreement on playing non-Nash equilibrium strategies (even non-correlated equilibria) of an *underlying game* may alter the incentives so that players will play according to the agreement. We also show that there is a conflict between the efficiency of the agreement and the incentives to respect it in symmetric games with strict strategic substitutes (Bulow, Roberts and Klemperer, 1985). On the other hand, in an important class of symmetric games where actions

are strategic complements such a conflict is circumvented: a symmetric efficient agreement can be made, if any.

Public good experiments with communication lend strong support for our theoretical finding: Isaac and Walker (1988) adopt a constant-returns-to-scale technology 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, McCue and Plott (1985) in a design with decreasing returns to scale.[1] Moreover, 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 (2006) 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.

We consider bilateral agreements in a wide array of strategic two-player interactions. The *underlying game*, whose strategies are being agreed upon, can be any normal form game. We abstract from how an agreement is established but assume that the *agreement* is either an action profile of the underlying game or disagreement. Given a game and players' proneness to guilt, each agreement maps the game into another game with the same strategy sets, but different payoffs. We are interested in which action profiles are *agreeable*. Agreeability is defined in terms of incentive compatibility and individual rationality. An action profile is *incentive compatible* if neither player prefers breaching. That is, for any unilateral deviation from the profile, the guilt cost is larger than or equal to the underlying game benefit for the deviator. We call the difference between the underlying game benefit and the guilt cost the *incentive to breach*.

If the agreement is established by pre-play negotiations, it is natural to think that each player can veto any agreement. We say that an action profile is *individually rational* if it ensures that each player gets more than in her least preferred Nash equilibrium of the underlying game.

Which agreements are agreeable will depend crucially on the properties of the guilt cost. We adopt the following properties:

{A} Guilt costs are weakly increasing in the harm that a player inflicts on her opponent by breaching an agreement.

{B} If the opponent breaches, then there is no guilt cost.

{C} Guilt costs are weakly increasing in the player's agreed payoff.

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

{D} If no agreement is reached, there is no guilt cost.

In addition to their intuitive appeal, there is experimental evidence that supports these assumptions in social psychology and in economics (see section 2). Property {A} captures the idea that if my breaching the agreement causes my opponent to lose a toe, I do not suffer more than if my breaching the agreement causes my opponent to lose a leg. Gneezy (2005) finds strong support for property {A}: his experiments suggest that people trade off the benefits of lying against the harm that lying inflicts on the opponent. Property {B} is a no-sucker property: I will not feel guilty about breaching an agreement if my opponent breaches the agreement, too. According to property {C} a kinder agreement's induces stronger guilt. To understand this property, notice that by property {B}, there is guilt only if the opponent does not breach the agreement. If the opponent respects and moreover the payoff is high if I respect, too, then the opponent is not only keeping his promise but he is also generous. Property {C} says that breaching the agreement and not reciprocating will induce stronger guilt than if the agreement had been less generous. Properties {B} and {C} render guilt reciprocal. Property {D} says that if there is no agreement about how the game should be played then there is no guilt.

Crucial for our finding in games with strategic complements and substitutes and an interesting result in its own right is that, in games where actions are ordered and the payoff is concave in each action, *checking that a marginal deviation from the agreement does not pay off is necessary and sufficient for incentive compatibility.*

Further towards our main conclusion, we find unambiguous effects on the marginal incentive to breach when the terms of the agreement are altered: *in symmetric games with strategic complements, changing either agreed action so as to improve a player's agreed payoff decreases her marginal incentive to breach.* These effects are quite natural and intuitive: if the terms of the agreement are better for me, I have a lower incentive to breach. Yet, the result does not hold generally: in *symmetric games with strategic substitutes*, *when the opponent's action is changed so as to improve the player's payoff, the marginal benefit increases and the marginal harm on the opponent decreases.* This is the source of our result, identifying a conflict between efficiency and incentives in symmetric games with strategic substitutes but no conflict in games with strategic complements.

The paper is organised as follows. Section 2 presents related literature in economics and psychology. Section 3 presents the model. Section 4 studies public good games. Section 5 presents general results and section 6 studies games with ordered strategy spaces. Section 7 concludes and discusses some further research problems.

2 Related literature

Economics. Evidence from experiments in public good games shows that even without communication subjects contribute positive amounts when contributing

nothing is a strictly dominant strategy in purely monetary incentives. Existing models of other-regarding preferences nicely capture this effect (Rabin, 1993; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000): people want to be fair to others that are contributing and contribute, too. Yet, a largely unexplained finding is that *communication* raises the contributions well above the amounts observed without communication (Ledyard, 1995). The earliest experiments show this in the prisoner's dilemma (Loomis, 1959; Radlow and Weidner, 1966). Recent studies for the two-person prisoner's dilemma case are provided by Duffy and Feltowich (2002) and (2006).[2]

A way forward in order to explain the effect of communication would be to combine one of the inequity aversion theories (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) with Farrell's (1987) idea that agreements will be stuck by if there is no incentive not to do so. With fairness preferences, it does not pay off to contribute little if others contribute a lot and thus, if players agree to contribute substantially, there is no incentive not to contribute. Yet, this fusion of theories is not completely convincing. First, it can only account for the experimental findings as long as the payoff functions are not too asymmetric, since if they are, identical contributions lead to unequal or unfair payoffs and players with payoffs below the average prefer trying to equate payoffs by breaching. Second, even in symmetric environments, if the more efficient symmetric equilibria exist in the underlying game, the learning process never reaches these equilibria in laboratory experiments when communication is not present and, yet, these outcomes are reached when communication is allowed for (Isaac and Walker, 1988; Isaac, McCue and Plott, 1985). Third, Ellingsen and Johanneson (2004) and Gneezy (2005) carry out further communication experiments and find behavioural patterns that cannot be explained by inequity aversion theories alone but which point to a preference for not lying.

The extensive form extension of Rabin's (1993) theory of reciprocity as introduced by Dufwenberg and Kirchsteiger (2004) is another candidate for capturing the phenomenon. Nevertheless, Charness and Dufwenberg (2003) show that sequential reciprocity can not fully account for the detected behavioural patterns related to communication. They conclude that there must be a separate preference related to lying and introduce, independently of the contribution of this paper (Charness and Dufwenberg, 2006), the guilt-aversion equilibrium, where a player suffers a cost when she acts counter to the opponent's expectation about her behaviour. Thus, like the theories of reciprocity, the theory falls into the category of psychological game theory (Geanokoplos, Pierce and Stachetti, 1989) where players' payoffs depend on beliefs explicitly (see also Dufwenberg, 2002; Battigalli and Dufwenberg, 2006).

The guilt-aversion theory is closely related to our approach. In the guilt-aversion model, promising to carry out an action is assumed to strengthen the belief that the opponent expects corresponding behaviour. Thereby the promise creates further incentives to behave accordingly. Notice however that the role of

[2]Extensions to public good provision games have been considered and the robustness of this result is verified by various experiments, for instance, Dawes, McTavish, and Shaklee (1977), Isaac, McCue, and Plott (1985), and Isaac and Walker (1988).

communication is only implicit in that model and it hinges on the assumption that beliefs are correlated to promises. In our model, preferences are defined over events that are observable in the laboratory - the agreements made and the actions chosen. Thus, with a suitable parameterisation, our model can be considered as a tractable model (Cox at al, 2006) of guilt-aversion. In addition to tractability and unlike in Charness and Dufwenberg (2006), in our approach a player may be more averse to let down a *justified* expectation or agreement but less averse to let down an *unjustified* expectation or agreement.[3] The precise meaning of justification in our approach derives from properties {B} and {C}: an opponent may feel guiltier about letting down the expectation of a player who respects an agreement and who is generous than a player who breaches or who only agrees to give very little. So as to property {D}, it seems crucial for the success of communication to foster cooperation that players have explicitly agreed on how to play and that this agreement is commonly known. Lev-on (2005) reviews communication experiments in public good games and concludes that mere identification or discussion which lacks explicit promising loses some of its effectiveness in supporting cooperation.[4]

Guilt has been discussed in several papers since Frank (1988) who argues that it may well be materially profitable for an agent to have a conscience - a dislike for disobeying social norms. Elster (1989) provides an extensive discussion about the role of social norms in economic theory. A recent model on emotional cost of breaching social norms is provided by Huck, Kübler, and Weibull (2003). These approaches involve no communication. Ellingsen and Johanneson (2004) are the first to propose a model of communication and guilt. They study the interplay of inequity aversion and guilt in a specific hold-up problem between a seller and a buyer. Their model is similar to ours in that guilt does not depend on the beliefs explicitly. In their model also, guilt is suffered if one breaches a promise. However, their model does not take into account the reciprocal elements of opponent's behaviour and it assumes that breaching a promise inflicts a constant guilt cost. This latter implies that their approach cannot account for the differences in behavior in public good games with constant and decreasing returns to scale mentioned in the introduction.

Psychology. In addition to their intuitive appeal, properties {A} to {D} are supported by experimental evidence and by psychological theory. As to property {A}, Hoffman (1982) suggests that guilt has its roots in a distress response to the suffering of others. The main empirical finding of Gneezy (2005) is that 1) lying is directly costly and 2) people do not care only about their own gain from lying: they are also sensitive to the harm that lying may inflict on others.

As far as property {B} is concerned, Baumeister, Stillwell, and Heatherton (1995) find that people feel guiltier about transgressions involving an "esteemed" person than about transgressions involving someone they hold in low regard. We implicitly assume that the esteem of a player towards a breaching opponent is smaller than if the opponent respects and, more specifically, we assume that the

[3] The relationship between guilt-aversion equilibrium and our approach is more extensively discussed in Miettinen (2006b).

[4] See also Brosig, Ockenfels and Weimann (2003).

player does not suffer at all from guilt if the opponent breaches the agreement.

Property {C} operates together with property {B}: the agreements that are respected and give a high payoff to a player signal the opponent's concern for the player's welfare and such opponents are likely to be esteemed. According to Clark and Mills (1984) and Clark (1979), the concern for the other's welfare is the defining feature of communal relationships as opposed to exchange relationships. According to Baumeister, Stillwell, and Heatherton (1995), guilt is more likely to arise in the former type than in the latter type of relationships.

So as to property {D}, an agreement explicitly states an expectation and a standard of behaviour for the play phase. Not reaching an agreement indicates players' inability to establish such a standard of a shared expectation. Millar and Tesser (1988) note that guilt depends on concurrence of one's own expectations of behaviour and those of the other person. Guilt appears mainly when there is a match in expectations of behaviour. Such a match of expectations is established by a pre-play agreement on how to play. On the other hand, some experimental studies on the public good game show that a single message for not contributing is sufficient to make an agreement invalid.[5] This body of research suggests each player should be able to veto an agreement and that if there is no agreement in place, guilt should be lower. We take this to an extreme and assume that there is guilt only if there is an agreement.

3 The model

Let Γ be a two-player simultaneous move normal form game, below referred to as the *underlying game*.[6] Before the game is played, an agreement is established. We abstract from whether this is done by a social norm or by pre-play negotiations, let alone which protocol is used, if any, if players negotiate.

We rule out the use of mixed strategies in the underlying game in order to simplify notation and to avoid taking a stand in an unresolved psychological question whether guilt is a function of consequences only or whether guilt is felt even if a mixed strategy different from the agreed one is chosen but the random draw picks up a pure strategy that is in the support of the agreed mixed strategy. Technically, an extension to mixed strategies could be easily done. We could also allow for agreements that condition the agreed actions on the outcomes of pre-game joint lotteries (Aumann, 1974).

3.1 The underlying game

The two-player *underlying game* is given by a normal form game $\Gamma = \{S_i,\ u_i(s) : S \to R, i = 1, 2\}$. The action set of player i is S_i. A combination of actions is an *action profile* $s = (s_i, s_j) \in S = S_i \times S_j$. The *underlying game payoff* of player i is $u_i(s)$. Notice that this payoff may well include social preference terms.

[5] See Ledyard (1995).

[6] The approach allows for a straightforward extension to sequential two stage games.

The lowest Nash payoff of player i is defined by $u_i^* \doteq \min_{s \in NE(\Gamma)} u_i(s)$ where $NE(\Gamma)$ is the set of pure Nash equilibria in the underlying game. The vector of such payoffs is $u^* = (u_i^*, u_j^*)$. If rational players play without pre-play negotiations and they have correct expectations about the behaviour of the other, then a Nash equilibrium should result. Thus, the lowest Nash payoff is the worst case scenario if the negotiations fail.

The negotiations or the convention establishes an agreement, m, on how to play, or disagreement, d. Thus, $m \in S \cup \{d\}$. If $m \in S$ is the agreement, then m_1 and m_2 are the *agreed actions* of players one and two respectively. The *agreed payoff* indicates how much more than u_i^* the player gets if both respect the agreement, $v_i(m) \doteq u_i(m) - u_i^*$. If player i deviates from the agreement, we get the *harm* on j by subtracting j's payoff at the deviation profile from his payoff at the agreed action profile, $h_j(m, s_i) \doteq u_j(m) - u_j(m_j, s_i)$. Similarly, i's *benefit from breaching* is $b_i(m, s_i) \doteq u_i(s_i, m_j) - u_i(m)$.

3.2 The entire game

Players are prone to guilt. If there is an agreement in place, they feel bad about not doing their part of the deal. Player i's *guilt cost*, $\theta_i g(v_i(m), h_j(m, s_i))$, depends on the agreed payoff and on the inflicted harm. The utility function in the entire game is assumed to be additively separable in guilt and the underlying game payoff.

$$U_i(m, s) = \begin{array}{l} u_i(s) - \theta_i g(v_i(m), h_j(m, s_i)) \text{ if } s_i \neq m_i \ , \ s_j = m_j \\ u_i(s) \text{ otherwise} \end{array} \qquad \text{(BD)}$$

The entire game payoff now depends on m and, due to guilt, talk is not cheap. The guilt cost is assumed to be non-negative. This rules out revengeful feelings or spite, on the one hand, and positive emotions related to respecting agreements, on the other hand. This is somewhat restrictive, but we wish to focus on guilt.

The parameters $\theta = (\theta_1, \theta_2)$ capture the players' *proneness to guilt*. For a given deviation, a player with higher proneness to guilt suffers more. We only allow for non-negative proneness to guilt, $\theta_i \in [0, \infty)$. If it is common knowledge that the proneness to guilt of both players equals zero, the model coupled with a communication protocol is one of cheap talk.[7]

Notice first, that the guilt cost depends on the agreement and on the deviation only indirectly through the agreed payoff and the harm. Second, choosing the agreed action m_i minimises the guilt cost at the second stage. Furthermore, (BD) implies that if there is disagreement about how to play, then there is no guilt cost. Also, there are no bad feelings about own cheating if the opponent cheats too. Thus (BD) introduces properties {B} and {D} into the guilt cost.

Moreover, we assume that the guilt cost is weakly increasing in the agreed

[7] As in Farrell (1987) but with any finite extensive form communication protocol ending up in an agreement.

payoff and in the harm. This is how we introduce properties {A} and {C}

$$g(v_i, h_j) \text{ is weakly increasing in } v_i \text{ and in } h_j \tag{AC}$$

Thus, if the guilt function is differentiable, then $\frac{\partial g}{\partial v_i} \geq 0$ and $\frac{\partial g}{\partial h_j} \geq 0$.

Also, we assume that if the player inflicts no harm on the opponent or if the agreed payoff equals the worst Nash payoff, then there is no guilt cost. Yet, we assume that if strictly positive harm is inflicted and the agreed payoff is strictly positive[8], then the guilt cost is strictly positive:

$$\begin{array}{c} g(v_i, h_j) > 0 \text{ if } h_j > 0,\ v_i > 0 \\ g(v_i, h_j) = 0 \text{ if } h_j = 0 \text{ or } v_i = 0 \end{array} \tag{EF}$$

Notice that these assumptions allow for a number of possible cost functions. For instance, a fixed guilt cost

$$g(v_i, h_j) = \begin{cases} \gamma \text{ if } h_j > 0,\ v_i > 0 \\ 0 \text{ otherwise} \end{cases}, \tag{1}$$

or a guilt cost that only depends on one of the arguments is allowed for. Another example of a guilt cost function with all the properties assumed in this section is

$$g(v_i(m), h_j(m, s_i)) = \max\{v_i(m), 0\}^{\gamma} \max\{h_j(m, s_i), 0\}^{\varphi}. \tag{2}$$

The entire game preferences of this form with $\gamma = \varphi = 1$ belong to the class of Cox-Friedman-Gjerstad (2006) preferences with the emotional state depending on the agreed payoff $v_i(m)$.

We suppose that the proneness to guilt types and the language are common knowledge. Thus, the players have correct point predictions about their opponent's proneness to guilt and the beliefs of all degrees coincide. Also, the players do not have to worry that the opponent might interpret an agreement to 'meet at noon' as an agreement to 'meet at quarter past noon.' Both these considerations are relevant but at this first step we abstract from this.[9]

Let us now introduce some further notation. Denote the underlying game best-reply correspondence of player i by $BR_i(s_j)$. By $\Gamma(m; \theta)$, denote the game where m is agreed and where the players' proneness to guilt is given by θ. Denote the equilibrium correspondence in that game by $s^*(m; \theta) = (s_i^*(m; \theta), s_j^*(m; \theta))$.

Let us write the payoffs of player i and player j respectively when player i deviates to s_i and player j respects the agreement, $s_j = m_j$, as

$$U_i(m_i, m_j, s_i, m_j) = u_i(m) + b_i(m, s_i) - \theta_i g(v_i(m), h_j(m, s_i)) \tag{3}$$

and

$$U_j(m_j, m_i, m_j, s_i) = u_j(m) - h_j(m, s_i). \tag{4}$$

[8] Our main results would be unaltered if we supposed that the reference point in the agreed payoff is the player's worst Pareto-efficient Nash payoff rather than the worst Nash payoff. This former is the lower bound for a long pre-play negotiation payoff derived in Rabin (1994), for instance.

[9] Notice also that since guilt depends on the agreement only indirectly, any permutation of the meanings of the agreements leaves the guilt unaltered.

where the first two entries of $U_i(.,.,.,.)$ are the agreed actions and the last two entries are the played actions of i and j respectively. These expressions give the players' entire game payoffs in terms of the agreed payoff, the benefit from breaching and the harm inflicted on the other when i breaches but not j. Player i's *incentive to breach* an agreement m is the difference between the benefit from breaching and the guilt cost, $B_i(m, s_i; \theta_i) \equiv b_i(m, s_i) - \theta_i g(v_i(m), h_j(m, s_i))$.

An agreement m is called *incentive compatible* if neither benefits from a unilateral deviation from the agreement,

$$\text{for all } s_i \in S_i \quad B_i(m, s_i; \theta_i) \leq 0 \tag{IC_i}$$

When this incentive compatibility condition holds for both players, the agreement m is a Nash equilibrium of the game where m is agreed upon, $\Gamma(m; \theta)$.

In the pre-play negotiations interpretation, we assume that each player can unilaterally enforce disagreement, d. In this case, an agreement m is called *individually rational* if no player prefers enforcing disagreement over playing m, i.e. if for $i = 1, 2$

$$u_i(m) \geq u_i^*. \tag{IR_i}$$

Here, the threat to the player who enforces d is the lowest payoff Nash equilibrium, u_i^*.

We now define *player i's potential to agree* as $A_i(\Gamma, \theta_i) \equiv \{m|\ m \text{ satisfies } (IC_i) \text{ and } (IR_i)\}$ and *the agreeable set* is defined as the intersection of the two potentials to agree, $A(\Gamma, \theta) \equiv \cap_{i=1,2} A_i(\Gamma, \theta_i)$. We call an action profile in i′s potential to agree *agreeable for i* and we call an action profile in the agreeable set simply *agreeable*.

4 Public good games

The prisoner's dilemma is a stylised version of a public good game. In the prisoner's dilemma there are two players who decide whether to contribute to the production of the public good or not. It is efficient that both contribute but it is a strictly dominant strategy not to contribute. We consider a prisoner's dilemma with the following payoffs:

	C	N
C	u_1, u_2	$u_1 - h_1, u_2 + b_2$
N	$u_1 + b_1, u_2 - h_2$	$0, 0$

(Prisoner's dilemma)

where $h_i > u_i > 0$ and $b_i > 0$ for $i = 1, 2$. Let us suppose that the guilt cost takes the simple form of (2) with $\gamma = \varphi = 1$. With this specification, player i respects an agreement to contribute, $m = (C, C)$, (given that the opponent does) if and only if

$$\theta_i \geq \frac{b_i}{u_i h_j} \tag{5}$$

An agreement on cooperation satisfying (5) is incentive compatible. Moreover, both contributing is individually rational by the structure of the prisoner's

dilemma. So, an agreement on (C, C) should be particularly easy to reach if b_i is small and h_j is large. Also, a large u_i facilitates cooperative agreements. This gives us comparative statics results that are testable.

In the prisoner's dilemma at outcomes (C, N) and (N, C), the payoff of one of the players is below the zero payoff in the equilibrium. Thus, these agreements are not individually rational. Both players not contributing, (N, N), is incentive compatible and individually rational for all types since it is the unique Nash equilibrium. Hence, (N, N) is always agreeable and (C, C) is agreeable if (5) holds for both players.

Moreover, notice that the individual rationality condition is actually redundant. It is implied by incentive compatibility: if individual rationality is violated, the agreed payoff falls below zero and guilt is zero. Thus any deviation which is beneficial in terms of the underlying game payoff will be made. It is easy to see that this property applies more generally. We will return to this issue in section 5.

In our model, proneness to guilt may transform a prisoner's dilemma into a coordination game. This is a familiar property of fairness models (Rabin, 1993; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). Yet here, first, the transformation is explicit and by agreement; and second, the ability to commit to contribute does not depend on how much more or less the opponent gets when the players cooperate, $u_i - u_j$. It depends on how much more the player gets when the players contribute than when they do not, $u_i - 0$. On the other hand, the player's payoffs are other-regarding only to the extent how much a player's defection affects the opponent's payoff.

We can easily generalise the prisoner's dilemma type of argumentation to public good games. Each player has an endowment of ten dollars. Each player decides how many dollars to contribute, $s_i \in \{0, ..., 10\}$. The payoff of player i reads

$$u_i(s) = G(\sum_{k=1,2} s_k) + 10 - s_i$$

where the production technology $G(.)$ maps the sum of contributions into the produced amount of the public good. We suppose that for all strategy profiles (s_1, s_2), $G'(\sum s) < 1$ where G' is the *marginal per capita return* (MPCR). Hence, it is a strictly dominant strategy, and thus a Nash equilibrium strategy, to contribute nothing. Whenever the *marginal group return* equals $2G' > 1$, it is socially optimal to increase one's contribution. In our model, we have two players only and the generalisation of the model for $n > 2$ players is left for future work.

Let us suppose in the rest of this section that the guilt cost is weakly convex in the harm on the other, $\varphi \geq 1$ and let the production technology have constant or decreasing returns to scale, $G'' \leq 0$. Players can agree to any agreement where both get a positive payoff and the guilt is sufficient to prevent either from breaching. As in the prisoner's dilemma, the individual rationality condition is implied by the incentive compatibility condition.

Notice further that due to the concavity of the UG payoff function in each

action, it is sufficient to check for the one dollar underprovision only: the benefit from breaching is concave and the harm on the other is convex as a rescaled negative of the opponent's payoff. Let us call the differential between the marginal benefit from breaching and the marginal guilt cost player i's *marginal incentive to breach*,

$$1 - G(\sum_{k=1,2} m_k) + G(\sum_{k=1,2} m_k - 1) \tag{6}$$
$$-\theta_i \max\{u_i(m), 0\}^{\gamma}[G(\sum m_k) - G(\sum m_k - 1)]^{\varphi}.$$

Supposing that an indifferent player respects the agreement, a player will breach if and only if (6) is positive. These marginal incentive compatibility conditions imply the incentive compatibility conditions which in turn imply the individual rationality conditions. Thus, the marginal incentive compatibility conditions are necessary and sufficient for agreeability (6).

A property explicit in (6) is worth emphasizing: if $G'' < 0$, there is a conflict between the efficiency of the agreement and the incentives to respect. To see this, notice that the harm on j due to a unit underprovision by i reads $h_j(m, m_i - 1) = G(\sum m_k) - G(\sum m_k - 1)$ which is decreasing in the sum of contributions. The marginal benefit for i from her unit underprovision vis-à-vis the agreement is $1 - h_j(m, m_i - 1)$. This is increasing in the sum of contributions. Since efficiency increases in the sum of contributions but the marginal harm on others decreases and the marginal benefit from breaching increases in the sum of contributions, the conflict is evident:

Proposition 1 *Let $G'' < 0$. If $k > 0$ $b_i(m_i, m_j, m_i - 1) < b_i(m_i + k, m_j + k, m_i + k - 1)$ and $h_i(m_i, m_j, m_j - 1) > h_i(m_i + k, m_j + k, m_j + k - 1)$ for $i = 1, 2$ $j \neq i$.*

If $\gamma > 0$, there is a countervailing force opposite to these marginal benefit and harm effects reported in proposition 1. Since by assumption efficiency is increased, so is the agreed payoff. This effect on guilt is positive, by {C}, and it tends to decrease i's incentive to breach.

Thus, whether or not incentive to breach increases in the sum of contributions depends on G'' and on φ, on the one hand (effect on the marginal benefit and harm), and on G' and γ, on the other hand (effect on the agreed payoff). If G'' is close to zero, the trade-off of the marginal benefit and the marginal harm is unaffected but the agreed payoff effect decreases the incentives to breach. Yet if G'' is substantially below zero and the agreed payoff does not much affect guilt, the effect of trading off the benefit and the harm increases the incentives to breach. Furthermore, if G'' is negative, the agreed payoff effect tends to fade away with efficiency. Eventually, if we have an interior group optimum, there will be a conflict between efficiency and the incentives as we are sufficiently close to the group optimum.

Yet, as a special case, if there are constant returns to scale, $G' = \alpha$, the marginal payoffs are constant and (6) reduces to

$$(1 - \alpha) - \theta_i (u_i(m))^{\gamma} \alpha^{\varphi}.$$

Thus, the changes in the agreed actions affect the breaching incentives only through the agreed payoff: the incentives to breach decrease in efficiency or at least they will not increase. Thus with constant returns to scale, the incentives and efficiency are aligned: if some disequilibrium strategy profile is agreeable, then an efficient profile is. Proposition 2 reports this finding and proposition 3 collects the rest of the findings of this section.

Proposition 2 *Let G' be constant. If s is such that for $i = 1, 2$ s is agreeable and $s_i > 0$, then there is an efficient action profile which is agreeable.*

Proposition 3 *In the public good game,*

(a) *an agreement is agreeable iff the marginal incentive to breach is non-positive for $i = 1, 2$.*

(b) *player i's marginal incentive to breach is increasing in m_i.*

(c) *if $G' = \alpha$, player i's marginal incentive to breach is decreasing in α and in m_j and in $\sum_{k=1,2} m_k$.*

(d) *if $G'' < 0$ and $\gamma = 0$, player i's marginal incentive to breach is increasing in m_j and in $\sum_{k=1,2} m_k$.*

Proof. To prove bullet (a), it is straightforward that

$$m \text{ satisfies } (IC_i) \text{ for } i = 1, 2 \Leftrightarrow m \text{ is agreeable,}$$

since (IC_i) implies (IR_i). It is easy to see that an upward deviation never pays off. Thus, it suffices to show that a non-positive marginal incentive to breach is equivalent to a non-positive incentive for deviating to any $s_i < m_i$. We have for all $s_i < m_i$

$$\begin{aligned} & m_i - G(\sum_{k=1,2} m_k) - s_i + G(m_j + m_i - s_i) \\ & -\theta_i(v_i(m))^\gamma (G(\sum_{k=1,2} m_k) - G(m_j + m_i - s_i))^\varphi \\ \leq \; & [1 - G(\sum_{k=1,2} m_k) + G(m_j + m_i - 1)](m_i - s_i) \\ & -\theta_i(v_i(m))^\gamma (G(\sum_{k=1,2} m_k) - G(m_j + m_i - 1_i))^\varphi (m_i - s_i) \end{aligned}$$

This follows from the fact that the opponent's payoff is increasing in s_i and that the guilt cost is convex in h_j. On the other hand,

$$\begin{aligned} & [1 - G(\sum_{k=1,2} m_k) + G(m_j + m_i - 1)](m_i - s_i) \\ & -\theta_i(v_i(m))^\gamma (G(\sum_{k=1,2} m_k) - G(m_j + m_i - 1_i))^\varphi (m_i - s_i) \\ \leq \; & 0 \end{aligned}$$

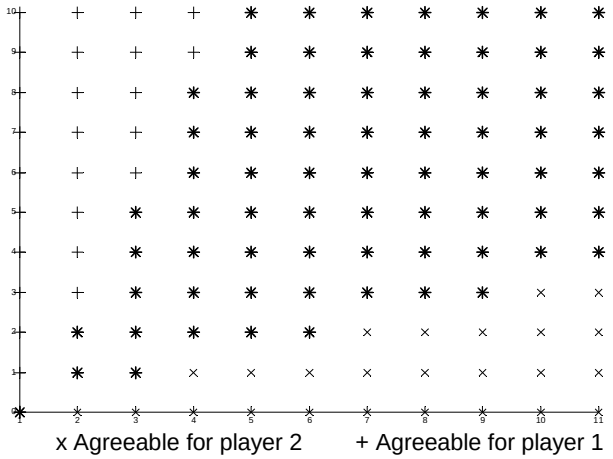

Figure 1: Figure 1: The agreeable set

$$\Leftrightarrow$$

$$\begin{aligned} & 1 - G(\sum_{k=1,2} m_k) + G(m_j + m_i - 1) \\ & -\theta_i(v_i(m))^{\gamma}(G(\sum_{k=1,2} m_k) - G(m_j + m_i - 1))^{\varphi} \\ \leq \quad & 0 \end{aligned}$$

since $m_i - s_i \geq 1$.

The proofs of bullets (b) to (d) rely on the effects of (b) m_i, (c) of α, m_j and $\sum_{k=1,2} m_k$ (when $G' = \alpha$) and (d) of m_j and $\sum_{k=1,2} m_k$ (when $G'' < 0$ and $\gamma = 0$) on (1) $b_i(m, m_i - 1)$, (2) on $u_i(m)$ and thus on $v_i(m)$ and (3) on $h_j(m, m_i - 1)$ when keeping in mind that the marginal incentive to breach increases in $b_i(m, m_i - 1)$ and decreases in $v_i(m)$ and in $h_j(m, m_i - 1)$. ■

Proposition 3 establishes that instead of checking for all possible deviations it is necessary and sufficient simply to check for a local deviation. Moreover, if $G' = \alpha$, then the marginal incentive to breach is monotone in each agreed action. Thus, to determine a player's potential to agree, we can look for agreements where the player is indifferent between respecting and deviating marginally. Any agreement where a player's action is smaller or an opponent's action is larger than at the boundary is agreeable for that player. Figure 1 shows the agreeable set for $G' = \alpha = \frac{3}{4}$, $\theta_i = 4$ with $\gamma = \varphi = 1$.

The action profiles that belong to player one's potential to agree are marked with plus signs and the action profiles that belong to player two's potential to agree are marked with crosses. Thus the action profiles marked with asterisks are agreeable action profiles, $A(\Gamma_{PG}(\frac{3}{4}), (4, 4))$. Notice that the best replies lie

on the axes and that the agreements where a player best-replies are agreeable for her. Thus, the Nash equilibrium, $(0, 0)$, is agreeable. Notice also that some efficient action profiles are agreeable, for instance, the symmetric efficient action profile where both give a full contribution, $m = (10, 10)$.

This section has illustrated that when communication is allowed for in public good games and players are prone to guilt, players may agree to contribute positive amounts and guilt may provide the necessary incentives to commit to the agreement. This is in line with a substantial body of experimental evidence (Ledyard, 1995). Further in regard to the experiments by Isaac, Mccue and Plott (1985) and Isaac and Walker (1988), we have suggested that the reason for lower cooperation rates in interior group optimum designs may not be due to the difficulties for the players to identify and agree on an interior group optimum, as suggested by Isaac and Walker. Rather the incentives to respect agreements sufficiently close to an interior group optimum are very weak, whereas the incentives to respect are the strongest close to a boundary group optimum. Notice that, in order to account for this difference, it is crucial that the guilt cost is weakly convex in the harm on the other. A constant guilt cost of deviating, (1) (Ellingsen and Johanneson (2004)), for instance, implies that guilt is concave in the harm on the other overall. Therefore their approach cannot account for the differences in the interior and boundary group optimum experiments.

Section 6 generalises the sharp contrast between the constant returns to scale technology and the decreasing returns to scale technology in public good production: we shall show that there is a conflict between incentives and efficiency in symmetric games with strategic substitutes where the payoff is monotone in the opponent's action. Such a conflict tends to be absent in symmetric games with strategic complements where the payoff is monotone in the opponent's action.

5 Properties of the agreeable set

In this section, we derive some simple properties that apply to any normal form underlying game in order to illustrate the functioning of the model. First, any underlying game (UG) Nash equilibrium is agreeable. Second, a Nash equilibrium of the UG remains a Nash equilibrium of most games where an agreement is in place. Yet, if an agreement is such that a player can unilaterally deviate to an UG Nash equilibrium, then this UG Nash equilibrium may no longer be a Nash equilibrium when the agreement is made. Third, if a player can deviate from an agreement and thereby benefit both players, the action profile is not agreeable. Yet, any individually rational profile that does not satisfy this latter property can be agreed upon if proneness to guilt is sufficiently high. This characterises the largest possible agreeable set as opposed to the smallest such set - the set of UG Nash equilibria. All of the proofs of this section and the following section are in the appendix.

If a player deviates from the UG best-reply, the guilt cost will only add to the forgone UG payoff. Thus a player can agree to play an UG best-reply if and

only if the agreement is individually rational. The first part of the following lemma establishes this finding.

On the other hand in lemma 1, we also establish that if a player's agreed action is not an UG best-reply to the opponent's agreed action, then the agreement is agreeable for the player if and only if it is incentive compatible. This is because then the UG benefit from breaching is positive at least when deviating to the UG best-reply and, if individual rationality is violated, then the agreement treats the player so badly that the guilt cost is zero. Thus, guilt does not deter breaching to the UG best-reply and the agreement is not incentive compatible.

Lemma 1 *Let $m_i \in BR_i(m_j)$. Then $m \in A_i(\Gamma, \theta_i)$ iff (IR_i) holds.*
Let $m_i \notin BR_i(m_j)$. Then $m \in A_i(\Gamma, \theta_i)$ iff (IC_i) holds.

This lemma is useful for characterizing each player's potential to agree: on the best-reply curve, all individually rational agreements are agreeable. Off the best-reply curve, all incentive compatible agreements are agreeable. Thus, for non-UG-equilibrium agreements only incentive compatibility matters. On the other hand, lemma 1 enables us establish that an UG equilibrium is agreeable for any proneness to guilt types. By definition, any Nash equilibrium payoff is individually rational. Thus by the first part of lemma 1, any Nash equilibrium belongs to each player's potential to agree. Thus, a Nash equilibrium is agreeable.

Proposition 4 *If $m \in NE(\Gamma)$, then $m \in A(\Gamma, \theta)$.*

First, for zero proneness to guilt types, Nash equilibria are the only agreeable action profiles.[10] Second, guilt never reduces the menu of the agreements available to the players. To the contrary, the public good example shows that positive proneness to guilt can substantially increase the set of profiles that are agreeable.

Recall that we ruled out mixed strategies and yet the only UG equilibrium may be in mixed strategies. Thus, an agreeable profile may not exist. Notice yet that allowing for mixed strategies would ensure that an agreeable profile always exists since, by proposition 4, an UG Nash equilibrium is agreeable and with mixed strategies a Nash equilibrium always exists in finite games.

Yet, pre-play negotiations may create an equilibrium selection problem when there is an agreement in place and players are prone to guilt. For instance, when players agree on cooperation in the prisoner's dilemma, defection remains an equilibrium of the transformed game. This is because if both players defect, neither feels guilt and therefore payoffs involve only underlying game payoffs. This insight is easily generalised: it is straightforward that an underlying game equilibrium where neither respects the agreement, m, is an equilibrium of the game $\Gamma(m; \theta)$. This shows that even if m is a Nash equilibrium of $\Gamma(m; \theta)$, there may be other equilibria as well.

[10] Aumann (1990) argues that cheap talk is credible only for a subset of Nash equilibria.

Lemma 2 *If for* $i = 1, 2,\ m_i \neq s_i^*$ *and* $s^* \in NE(\Gamma)$ *then* $s^* \in NE(\Gamma(m;\theta))$

The equilibrium selection problem apparent in lemma (2) is avoided however if we suppose that players will conform to the agreement, if there is no incentive not to do so, as assumed in Farrell (1987).[11]

Even if communication with guilt may increase the number of equilibria, it may also remove an equilibrium from the game. Consider the following game of chicken:

	L	R
T	$0, 0$	$3, 1$
B	$1, 3$	$2, 2$

(7)

The Nash equilibria of this game are (B, L) and (T, R). Let us suppose that player one's proneness to guilt is two, $\theta_1 = 2$ and the guilt cost function is as in (1) with $\gamma = 1$. Let us suppose that players agree on playing (B, R) which gives an agreed payoff of 2 for player one. Now, if player one breaches the agreement and chooses T instead, she gets $3 - 2 = 1$ which is smaller than 2 and, thus, (T, R) is not an equilibrium when players have agreed on (B, R) even if it is a Nash equilibrium of the underlying game.

Next, notice that an agreement where one of the players can make both players better off by deviating unilaterally from the agreement (even if the opponent respects the agreement) does not belong to the agreeable set.

Lemma 3 *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* $m \notin A(\Gamma, \theta)$ *for any* θ.

Lemma 3 follows immediately from the monotonic payoffs (AC) and the strict cost (EF) conditions: when the harm inflicted on the other is non-positive, there is no guilt cost. Since a player can make herself better off, she will do so and the agreement is not incentive compatible.

Thus, for instance pattern (B, L) is never agreeable in the following game:

	L	R
T	$2, 2$	$0, 100$
B	$1, 1$	$1, 1$

(8)

since, if there is an agreement on (B, L) in place, player one breaches and chooses T, both players are better off. Perhaps player one does not breach (B, L) because she understands that then player two has an incentive to choose R which would make her worse off than in (B, L). However, player one would then be inclined to choose B again. Agreeing on (B, L) would thus leave a lot of room for rationalizing various kinds of play and truth is no more focal in the sense of Farrell (1987). Indeed, this type of ambiguity may also question whether (B, L) is agreeable in the first place.

[11] Applying Farrell (1987), we may refine the Nash equilibrium concept in the game $\Gamma(m;\theta)$ by assuming that if m is a Nash equilibrium of $\Gamma(m;\theta)$, then m will be played, $s^*(m;\theta) = m$.

In (8), players cannot agree on (T, R) either, since player 1 gets a smaller payoff than in the underlying game equilibrium, (B, R). On the other hand, if player 2's proneness to guilt is small, players will not be able to agree on (T, L) either due to player two's high gain from choosing R instead. Nevertheless, if we let player two's proneness to guilt become sufficiently high, (T, L) becomes agreeable. As the proneness to guilt becomes infinite, the guilt cost becomes infinite for deviations that cause positive harm. Hence, whenever deviation causes harm, it will not be made. In general, if UG payoffs are finite, with sufficiently high proneness to guilt all individually rational profiles are agreeable for which a Pareto-improving deviation does not exist, and no other profile is.

Proposition 5 *Let the underlying game payoffs be finite. Let* $v_i(m) > 0$ *for* $i = 1, 2$. *Then* $m \in \lim_{\theta_1 \to \infty, \theta_2 \to \infty} A(\Gamma, \theta)$ *iff for* $i = 1, 2$ *and for all* s_i, $u_i(m) \geq u_i(s_i, m_j)$ *or* $u_j(m_j, s_i) < u_i(m)$

If the set of Nash equilibria is the smallest set that is agreeable (cheap-talk), proposition 5 describes the largest possible agreeable set, the agreeable set for types that are infinitely prone to guilt.

Lemma 3 has another implication, which is mentioned without a proof. Namely, within the agreeable set, the interests of the players are opposed for any change in one of the agreed actions.

Corollary 1 *Let* $(m_i, m_j), (m_i', m_j) \in A(\Gamma, \theta)$ *then*

$$\begin{array}{l} u_i(m_i, m_j) > u_i(m_i', m_j) \Rightarrow u_j(m_i', m_j) > u_j(m_i, m_j) \\ u_j(m_i', m_j) > u_j(m_i, m_j) \Rightarrow u_i(m_i, m_j) \geq u_i(m_i', m_j) \end{array} \tag{9}$$

6 Finite games with ordered strategy spaces

Let us now focus on finite symmetric games with *ordered* strategy spaces. Without loss of generality we label the actions from 0 to n, $S_i = \{0, ..., n\}$. Inspired by the results in the public good production with decreasing and constant returns to scale, we generalise the results gained there: There is a conflict between the incentives and the efficiency in symmetric games with strategic substitutes where the payoffs are monotone in the opponent's action (decreasing returns to scale). Such a conflict tends to be absent in symmetric games with strategic complements and monotone payoffs in the opponent's action (constant returns to scale). Potters and Suetens (2006) find strong empirical support for these findings.

We now adopt some new concepts and notational simplifications. For $s \in S$ and for $k \in \mathbb{Z}$, we let $s + k \doteq (s_i + k, s_j + k)$. We let the marginal benefit from breaching be defined as $\beta_i(m_i, m_j) \doteq b_i(m_i, m_j, m_i - 1)$ and the marginal harm as $\eta_i(m_i, m_j) \doteq h_i(m_i, m_j, m_j - 1)$. Thus $\beta_i(m + k) = \beta_i(m_i + k, m_j + k)$, $\eta_i(m + k) = \eta_i(m_i + k, m_j + k)$ and $u_i(m + k) = u_i(m_i + k, m_j + k)$ for $k \in \mathbb{Z}$.

We first set the scene by making *further assumptions on the underlying game.*

{1} The payoff of player i is increasing in the action of player j.

{2} The player's payoff is concave in her own action and in that of the opponent. That is, for all s

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

and for all s

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

Also, we make a *further assumption on the guilt cost*:

{4} g is convex in h_j

Throughout this section, in addition to supposing that the game is finite and symmetric, we suppose that properties {1}, {2} and {4} hold. We wish to compare how the incentives to breach are affected by the efficiency of a symmetric agreement in games with strategic complements, on the one hand, and with strategic substitutes, on the other hand. Thus we make two alternative assumptions:

{3} The actions are strategic complements[12]. That is for all s

$$\phi_i(s) \doteq u_i(s_i, s_j) - u_i(s_i, s_j-1) - [u_i(s_i-1, s_j) - u_i(s_i-1, s_j-1)] \geq 0.$$

{3'} The actions are strategic substitutes, $\phi_i(s) < 0$.

Properties {1}, {2},{3} and {4} are satisfied in the linear public good game, $G' = \alpha$, in a degenerate manner: for all s, $\delta_i(s) = \sigma_i(s) = \phi_i(s) = 0$. On the other hand, {1}, {2},{3'} and {4} are satisfied in the public good game with decreasing returns to scale, $G'' < 0$, where each partial second differential is negative.

Notice that 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 since the harm is just a rescaled negative of the underlying game payoff, $h_j(m, s_i) \doteq u_j(m) - u_j(m_j, s_i)$. Thus, by assumption {4}, the guilt cost 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. Consequently, checking that neither prefers breaching the agreement marginally is necessary and sufficient for an agreement to be incentive compatible. Notice that assumption {4} rules out the constant cost of breaching, (1), since with that specification guilt is concave in harm.

To simply formulate such a marginal incentive condition we extend the concept of the marginal incentive to breach from the context of the public good game.

[12] See Bulow et al. (1985).

Definition 1 *(Marginal incentive to breach)*

$$\mathbb{B}_i(m, \theta_i) \doteq \beta_i(m) - \theta_i g(v_i(m), \eta_j(m))$$

When $u_i(m_i+1, m_j) - u_i(m) > 0$, player i gains in the UG payoff from choosing an action higher than the one agreed upon. Moreover due to assumption $\{1\}$, such a deviation does not make the opponent worse off and thus player i does not suffer from guilt in this case. Consequently, each player's agreed payoff must necessarily be non-increasing[13] at any action profile which is agreeable. We denote this set by M^F

$$M^F \doteq \cap_{i=1,2} M_i^F \text{ where } M_i^F \doteq \{m | \beta_i(m_i + 1, m_j) \geq 0\}. \tag{10}$$

$\mathbb{B}_i(m, \theta_i)$ characterises player i's marginal breaching incentive in the set M_i^F

Next, we establish a necessary and a sufficient condition for agreeability that generalises our finding in the public good game. Above we made the remark that, due to the convexity of the problem, there is no incentive to breach the agreement at the margin if and only if there is no incentive to breach at all. Second by lemma 1, incentive compatibility implies individual rationality when off the underlying game best-reply curves. Thus, we have the following:

Proposition 6 *Let $m_i \neq BR_i(m_j)$, $m_i \in M_i^F$ and $m_i \notin \{0, n\}$. An action profile is agreeable for i if and only if i's marginal incentive to breach is non-positive.*

In the public good game with $G' = \alpha$, we found that the marginal incentive to breach is monotone in each agreed action. This property holds more generally in games where actions are strategic complements, $\phi \geq 0$.

First notice that, due to the concavity of a player's payoff in each action, the marginal benefit from breaching increases and the marginal harm on the other decreases in the player's own action within the agreeable set. This is because the player's own UG payoff is decreasing more rapidly and the opponent's payoff is increasing less rapidly if one's own action is higher and because a player deviates down if at all in M_i^F.

On the other hand, due to the strategic complementarities, the marginal downward deviation pays worse off if the opponent's agreed action is higher; moreover in this case, more harm is inflicted on the opponent. Yet, in games with strict strategic substitutes, increasing the opponent's action has the reverse impact: the marginal benefit from breaching increases and the marginal harm on the other decreases. The following lemma summarises:

Lemma 4 $\beta_i(m_i + 1, m_j) - \beta_i(m_i, m_j) = -\delta_i(s)$
$\eta_j(m_i + 1, m_j) - \eta_j(m_i, m_j) = \sigma_j(s)$
$\beta_i(m_i, m_j + 1) - \beta_i(m_i, m_j) = -\phi_i(s_i, s_j + 1)$
$\eta_j(m_i, m_j + 1) - \eta_j(m_i, m_j) = \phi_j(s_j + 1, s_i)$

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

Notice that, conditional on both respecting, the agreed payoff reflects a player's preference ordering over the agreements. Lemma 4 shows that, in games with strategic complements, if an agreed action is changed so as to increase a player's agreed payoff, her incentives to breach decrease. This sounds natural: a player has a stronger incentive to respect a more generous agreement. The incentive to respect is further strengthened by the agreed payoff effect, {C}, entering through the marginal guilt cost. To be precise, for this latter claim to hold generally, we need to assume the following:

{5} g is convex in v_i and supermodular[14] in its arguments.

Assuming {5} in games with strategic complements, when m_i or m_j is changed so as to increase i's agreed payoff, there is an unambiguously negative impact on the marginal incentive to breach.

Proposition 7 *Let {3} and {5} hold. Then i's marginal incentive to breach is non-decreasing in m_i and non-increasing in m_j in the agreeable set.*

Yet, in symmetric games with strategic substitutes where $\phi < 0$, there is some conflict between the preference over the agreements and the incentive to respect them: when the opponent's agreed action is increased, the agreed payoff increases but the incentive to respect may decrease since the marginal benefit increases and the marginal harm decreases. This implies a conflict between the efficiency of an agreement and the incentives to respect it in games with strategic substitutes, because both agreed actions must be increased relative to an interior equilibrium to improve efficiency in any game satisfying the unifying assumptions of this section.

Theorem 1 *Let {3'} hold. Let s^* be an interior equilibrium. If $u_i(s^* + k) - u(s^*) > 0$ for $k \in \mathbb{Z}$ and for $i = 1, 2$ then $\beta_i(s^* + k) > 0$ and $\eta_i(s^* + k) < \eta_i(s^*)$ for $i = 1, 2$.*

However, there is often no such conflict in games with strategic complements. In theorem 2, we consider improving the efficiency from the most efficient interior UG equilibrium which we know to be agreeable, by proposition 4. We conclude that, if players cannot agree on a symmetric profile close to it but that they can agree on a more efficient action profile (third and sixth bullet), then the players can agree on a fully efficient profile.

Theorem 2 *Let {3} hold and let $\delta \leq 0$, $\sigma \leq 0$ and $\phi \geq 0$ be constants. Let s^* be the most efficient interior UG equilibrium. Let $s^* + \overline{k}$ be an efficient action profile.*

If

- *$u_i(s + k)$ is convex in k.*

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

- *g satisfies {5}.*
- *for each* $i = 1, 2$ *there are* k_i' *and* k_i'' *s.t.* $0 \leq k_i'' < k_i'$, $\mathbb{B}_i(s^* + k_i'', \theta_i) \geq 0$ *and* $\mathbb{B}_i(s^* + k_i', \theta_i) \leq 0$

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

- *the marginal harm,* $\eta_j(s + k)$*, is non-decreasing in* k
- *the guilt cost is unaffected by the agreed payoff,* $g(v', \eta) = g(v, \eta)$ *for all* η *and* $v', v > 0$
- *for each* $i = 1, 2$ *there are* k_i' *and* k_i'' *s.t.* $0 \leq k_i'' < k_i' < \overline{k}$, $\mathbb{B}_i(s^* + k_i'', \theta_i) \geq 0$ *and* $\mathbb{B}_i(s^* + k_i', \theta_i) \leq 0$.

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

Notice that we need some rather unrestrictive assumptions to establish the result. Either we need to assume that the UG payoff is convex when increasing both actions simultaneously while keeping them identical (first bullet) and that the harm and the agreed payoff enter in a convex and complementary manner in the guilt function (second bullet). Alternatively, we need to assume that the marginal harm does not decrease in such simultaneous changes of both actions (fourth bullet) and that guilt is unaffected by the agreed payoff (fifth bullet, this corresponds $\gamma = 0$ in (2)). Our assumptions guarantee that, as the efficiency of the agreement is increased, if the increasing marginal harm and the increasing agreed payoff balance out the UG incentive to breach the agreement, then this balance will hold for any agreement which is more efficient than this cut-off agreement.

If strategic complementarities are strong, then the action profile with the greatest feasible actions is the most efficient profile and, moreover, it is a Nash equilibrium of the UG. Therefore, by proposition 4, it is agreeable. This case is somewhat uninteresting for us since guilt plays no role in achieving efficiency. Pre-play negotiations or norms then enact as a mere coordination device when choosing among multiple equilibria. Our theorem establishes that, even in the more interesting case where actions are weaker strategic complements and the efficient profile is not an equilibrium, norms and pre-play negotiations tend to achieve first-best efficiency if any improvements to efficiency can be achieved at all.

An interesting corollary of the theorem concerns symmetric games with a unique equilibrium. In that case if any symmetric profile between the equilibrium and the efficient profile is agreeable, then an efficient profile is agreeable.

Corollary 2 *Let {3} hold and let* $\delta \leq 0$, $\sigma \leq 0$ *and* $\phi \geq 0$ *be constants. Let* s^* *be the unique UG equilibrium. Let* $\beta_i(s^*) = 0$ *for* $i = 1, 2$. *Let* $s^* + \overline{k}$ *be an efficient action profile.*
If

- $u_i(s+k)$ *is convex in* k
- g *satisfies {5}*
- *for each* $i = 1, 2$ *there is* k_i' *s.t.* $\mathbb{B}_i(s^* + k_i', \theta_i) \leq 0$.

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

- *the marginal harm,* $\eta_j(s+k)$*, is non-decreasing in* k
- $g(v', \eta) = g(v, \eta)$ *for all* η *and* $v', v > 0$
- *for each* $i = 1, 2$ *there is* k_i' *s.t.* $k_i' < \overline{k}$ *and* $\mathbb{B}_i(s^* + k_i', \theta_i) \leq 0$

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

Regarding theorems 1, 2 and corollary 2, notice again that assumption {1} was made without the loss of generality. All we need is symmetry. If the payoffs are decreasing in the opponent's action, we can restore assumption {1} by reversing the ordering of each strategy set. This has no effect on concavity or strategic complementarities. Thus, symmetric games with decreasing payoffs in the opponent's action can be analysed using the same artillery.

Above in section 4, we put forward some evidence in the public good provision context supporting theorems 1 and 2 and corollary 2. Moreover, Suetens (2005) finds in a laboratory experiment that communication induces cooperation in an R&D environment with strategic complements but no cooperation in an environment with strategic substitutes. Suetens and Potters (2006) review data from four duopoly experiments and find that there is more tacit collusion in Bertrand duopolies than in Cournot duopolies[15] with substitute products implying strategic complements and substitutes, respectively. Potters and Suetens (2006) come up with a clever design to extract the effect of strategic complementarities on cooperation in a general unframed laboratory experiment. In the games that they study, there are constant second order effects on the UG payoffs and a unique symmetric equilibrium which is also an interior equilibrium, in line with corollary 2. They find substantially more cooperation in games with strategic complements than in games with strategic substitutes. Thus, the present approach organises rather well the differences in the effects of communication in a substantial number of games and experiments.

7 Discussion

The main contribution of this paper is to provide a game theoretic approach to pre-play agreements by negotiations, conventions or social norms when people may feel guilty about breaching an agreement. The model incorporates the

[15] Miettinen (2006 a, b) present a Cournot duopoly with perfect substitutes as an example of a game where a conflict between efficiency and incentives to respect prevails.

most important stylised facts about guilt that research in social psychology and experimental economics has established.

We show that social norms and pre-play negotiations combined with guilt may have plausible effects on strategic interaction. Trivially, the set of agreeable outcomes may be larger than the set of underlying game Nash equilibria since the guilt cost provides an extra incentive to comply with an agreed action profile. Moreover, this effect may prevail even if monetary stakes are high provided that the harm that a defection inflicts on the opponent is sufficiently high. The approach is in line with results from public good experiments where communication significantly increases contribution levels (Ledyard, 1995). The result extends to a large class of games with a public good structure: team production, collusion in Bertrand and Cournot duopolies, R&D etc. Yet, there is an important distinction as to whether the approach predicts an efficient or an inefficient agreement: in symmetric games with strategic complements, if a symmetric non-underlying game equilibrium is agreeable, then a symmetric efficient action profile is agreeable. On the other hand, in symmetric games with strategic substitutes, there tends to be a conflict between the incentives to respect an agreement and its efficiency. Several experiments provide support for these findings.

Our result indicates that, when trying to promote efficiency in social interactions, at the workplace for instance, relying on norms and informal agreements may be a good idea if the interaction is characterised by strategic complements but that formal agreements may be needed when strategic substitutability holds.

This latter consideration does not concern only infrequent one-shot interactions. Very similar results would be obtained if we supposed that players have zero proneness to guilt and they informally agree on a stationary outcome in an infinitely repeated analog of the underlying game and the punishment paths are exogenously determined (in a social contract, for instance). If the agreement is breached, it takes some time to detect breaching, and players revert to punishment strategies for a length of time that depends on the deviator's agreed payoff and the harm she inflicts on the other.[16]

This paper has not analysed the effect of the negotiation protocol on the agreement. Another dimension for future research is the relaxation of the assumption of complete information on proneness to guilt types. The choice of an optimal agreement when information is private requires trading off the player's own agreed payoff with the probability that the opponent breaches the agreement. Notice yet that if the information on proneness to guilt is private, signalling seems not to be an issue in pre-play negotiations: the maximisation problem conditional on respecting is the same independently of the type and thus all types that intend to respect behave identically. Any type who intends to breach is thus detected.

On the other hand, a dynamic setup of incomplete information on proneness to guilt would allow for the players to build up reputations. First, it may be

[16] In this case, if the social contract implements punishments proportional to the violation ("tooth for tooth - eye for eye"), the folk theorem does not hold. See Miettinen (2006a) for further details.

optimal for types with high proneness to guilt to build up a reputation for lower proneness to guilt so that, in pre-play negotiations, they are proposed higher shares of the surplus in the future. Second, types with low proneness to guilt may be willing to build up a reputation for higher proneness to guilt in order to be able to reach agreements with a larger fraction of types. From a similar perspective, one can study the evolution of proneness to guilt for a given stochastic process of games and matches.

8 Appendix

8.1 Proof of lemma 1

$m_i \in BR_i(m_j) \Leftrightarrow$ for all s_i, $u_i(m_i, m_j) \geq u_i(s_i, m_j) \Rightarrow$ for all s_i, $u_i(m_i, m_j) \geq u_i(s_i, m_j) - g(v_i(m), h_j(m, s_j)) \Leftrightarrow$ for all s_i, $B_i(m, s_i; \theta_i) \leq 0 \Leftrightarrow (IC_i)$. Thus, $m \in A_i(\Gamma, \theta_i)$ iff (IR_i).

For the second claim, $m_i \notin BR_i(m_j) \Rightarrow$ there is s_i' such that $u_i(s_i', m_j) > u_i(m_i, m_j)$. Suppose now that (IC_i) holds. But, for all s_i, $B_i(m, s_i; \theta_i) \leq 0 \Rightarrow B_i(m, s_i'; \theta_i) \leq 0 \Rightarrow g(v_i(m), h_j(m, s_i')) \geq u_i(s_i', m_j) - u_i(m_i, m_j) \Rightarrow g(v_i(m), h_j(m, s_i')) > 0 \Rightarrow v_i(m) > 0$. Thus (IR_i) holds and $m \in A_i(\Gamma, \theta_i)$.

Suppose now that m is agreeable. Then by definition (IC_i) holds.■

8.2 Proof of proposition 4

Since m is an UG Nash equilibrium, $u_i(m) \geq \min_{s \in NE(\Gamma)} u_i(s)$ and thus $v_i(m) \geq 0$ for $i = 1, 2$. Since m is a Nash equilibrium in Γ, $m_i \in BR_i(m_j)$ for $i = 1, 2$. Then, by lemma (1), $m \in A_i(\Gamma, \theta_i)$ for $i = 1, 2$ and, by definition, $m \in A(\Gamma, \theta)$.■

8.3 Proof of lemma 2

Since both deviate from the agreement the guilt cost is zero for both. Then for all s_i, $U_i(m, s^*) = u_i(s^*) \geq u_i(s_i, s_j^*) = U_i(m, s_i, s_j^*)$ where the inequality follows from the fact that s^* is a Nash equilibrium of Γ.■

8.4 Proof of lemma 3

Conditions (AC) and (EF) imply that $g(v_i(m), h_j(m, s_i)) = 0$ if $h_j(m, s_i) < 0$. But indeed, $h_j(m, s_i) \doteq u_j(m) - u_j(m_j, s_i) < 0$ by assumption and also by assumption $b_i(m, s_i) = u_i(s_i, m_j) - u_i(m) > 0$. Thus for any θ_i, $B_i(m, s_i, \theta_i) = b_i(m, s_i) - \theta_i g(v_i(m), h_j(m, s_i)) > 0$. Therefore (IC_i) is violated and $m \notin A_i(\Gamma, \theta_i)$ and thus $m \notin A(\Gamma, \theta)$.■

8.5 Proof of proposition 5

By assumption, $v_i(m) > 0$ for $i = 1, 2$. Take player i and an arbitrary s_i. First, if $u_i(m_i, m_j) \geq u_i(s_i, m_j)$ then $u_i(m_i, m_j) \geq u_i(s_i, m_j) - g(v_i(m), h_j(m, s_j))$ and $B_i(m, s_i; \theta_i) \leq 0$. Second, if $u_j(m_j, s_i) < u_j(m_j, m_i)$ then $h_j(m, s_i) > 0$. By,

(EF) $g(v_i(m), h_j(m, s_i)) > 0$. Thus, since payoffs in Γ are finite, $\lim_{\theta_i \to \infty} \theta_i g(v_i(m), h_j(m, s_i)) \geq u_i(s_i, m_j) - u_i(m_i, m_j)$. Hence, $\lim_{\theta_i \to \infty} B(m, s_i; \theta_i) \leq 0$. Since either $u_i(m_i, m_j) \geq u_i(s_i, m_j)$ or $u_j(m_j, s_i) < u_j(m_j, m_i)$ holds for every s_i, (IC_i) holds. Thus $m \in A_i(\Gamma, \theta_i)$. This is true for both players. Thus, $m \in A(\Gamma, \theta)$.

Let now $m \in A(\Gamma, \theta)$. Suppose to the contrary that there is i and s_i such that neither $u_i(m_i, m_j) \geq u_i(s_i, m_j)$ nor $u_j(m_j, s_i) < u_j(m_j, m_i)$ holds. Therefore, by lemma 3, $m \notin A(\Gamma, \theta)$. This is a contradiction.■

8.6 Proof of proposition 6

Lemma 5 *Let Γ be finite. Let $m_i \neq \{0, n\}$, $m_i \in M_i^F$ and $m_i \notin BR_i(m_j)$. Let {1}, {2} and {4} hold. Then (IC_i) holds if and only if $\mathbb{B}_i(m, \theta_i) \leq 0$.*

Proof. We will show that (IC_i) does not hold iff $\mathbb{B}_i(m, \theta_i) > 0$.

Let $\mathbb{B}_i(m, \theta_i) > 0$. By the definition of $\mathbb{B}_i(m, \theta_i)$, $B(m, m_i - 1; \theta_i) = \beta_i(m) - \theta_i g(v_i(m), \eta_j(m)) > 0$ and (IC_i) is violated.

Let (IC_i) be violated. Thus, there is s_i' such that $B_i(m, s_i'; \theta_i) > 0$. Suppose to the contrary that $\mathbb{B}_i(m, \theta_i) \leq 0$ and thus

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

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 $m_i \in BR_i(m_j)$ 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 $u_i(m_i - 1, m_j) - u_i(m_i, m_j) = \beta_i(m) > 0$ and thus $g(v_i(m), h_j(m, m_i - 1)) > 0$ since $\mathbb{B}_i(m, \theta_i) \leq 0$. By assumption {1}, the harm increases in deviations further downwards. Also by assumption {4} guilt cost is convex in h_j and by assumption {2} 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 {2}, the payoff u_i is concave in s_i and thus the benefit from breaching $u_i(s_i, m_j) - u_i(m_i, m_j)$ is concave in s_i. Thus if $\mathbb{B}_i(m, \theta_i) \leq 0$ then $B(m, s; \theta_i) \leq 0$ for all $s_i < m_i$. We have a contradiction. ■

Proof of the proposition

The result follows directly from lemma 1, lemma 5 and the fact that $A(\Gamma, \theta) = \cap_{i=1,2} A_i(\Gamma, \theta_i)$■

8.7 Proof of lemma 4

$\beta_i(m_i + 1, m_j) - \beta_i(m_i, m_j) = u_i(m_i, m_j) - u_i(m_i + 1, m_j) - [u_i(m_i - 1, m_j) - u_i(m_i, m_j)] = -\delta_i(m)$

$\beta_i(m_i, m_j + 1) - \beta_i(m_i, m_j) = u_i(m_i - 1, m_j + 1) - u_i(m_i, m_j + 1) - [u_i(m_i - 1, m_j) - u_i(m_i, m_j)] = -\phi_i(m_i, m_j + 1)$

$\eta_j(m_j, m_i + 1) - \eta_j(m_j, m_i) = u_j(m_j, m_i + 1) - u_j(m_j, m_i) - [u_j(m_j, m_i) - u_j(m_j, m_i - 1)] = \sigma_j(m)$

$\eta_j(m_j + 1, m_i) - \eta_j(m_j, m_i) = u_j(m_j + 1, m_i) - u_j(m_j + 1, m_i - 1) - [u_j(m_j, m_i) - u_j(m_j, m_i - 1)] = \phi_j(m_j + 1, m_i)$■

8.8 Proof of proposition 7

The marginal incentive to breach reads

$$\mathbb{B}_i(m_i, m_j) = \beta_i(m_i, m_j) - \theta_i g(v_i(m_i, m_j), \eta_j(m_i, m_j))$$

But $\beta_i(m)$ is non-decreasing in m_i and $\eta_j(m)$ is non-increasing in m_i by lemma 4. Also, $u_i(m_i+1, m_j) - u_i(m_i, m_j) \leq 0$ since m is agreeable for i. This implies that $u_i(m)$ is non-increasing in m_i. But g is non-decreasing in both arguments. Thus, $\mathbb{B}_i(m_i, m_j)$ is non-decreasing in m_i.

On the other hand, $\beta_i(m_i, m_j)$ is non-increasing in m_j and $\eta_j(m_i, m_j)$ is non-decreasing in m_j by lemma 4. Also, u_i is increasing in m_j by assumption. But g is non-decreasing in both arguments. Thus, $\mathbb{B}_i(m_i, m_j)$ is non-increasing in m_i.■

8.9 Proof of theorem 1

Since the equilibrium is interior, by the concavity of the payoffs in own actions, we must have $\beta_i(s^*) \leq 0$ and $\beta_i(s_i^* + 1, s_j^*) \geq 0$ for $i = 1, 2$. Since {3'} holds, by lemma 4, $\beta_i(s_i^*, s_j^* - 1) < 0$ and therefore, by {1}, $u_i(s^*) - u_i(s^* - 1) = u_i(s^*) - u_i(s_i^*, s_j^* - 1) - \beta_i(s_i^*, s_j^* - 1) > 0$. More generally by lemma 4 and assumptions {3'} and {1}, $\beta_i(s_i^* + 1 - k, s_j^* - k) < 0$ 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 $k > 0$. But, since $\phi_i(s) < 0$ and $\delta_i(s), \sigma_i(s) \leq 0$, by lemma 4, $\beta_i(s^* + k) > \beta_i(s^*)$ and $\eta(s^* + k) < \eta(s^*)$ for $i = 1, 2$.■

8.10 Proof of theorem 2

In this subsection, we suppose throughout that ϕ is a non-negative constant and that δ and σ are non-positive constants and that Γ is symmetric and that {1} and {4} hold.

Lemma 6 *Let $-\delta > \phi$. Let s^* be the most efficient interior UG. Then*

- $\beta_i(s^*) \leq 0$
- $\beta_i(s_i^* + 1, s_j^*) \geq 0$
- *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
- $\beta_i(s_i^* + 1, s_j^* + 1) > 0$

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 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 $\beta_i(s_i^* + 1, s_j^* + 1) \leq 0$. Now $\beta_i(s_i^* + 1, s_j^*) \geq 0$ implies, by lemma 4, that $\beta_i(s_i^* + 2, s_j^* + 1) > 0$ since -$\delta > \phi$. By symmetry this holds also for j. Thus $s^* + 1$ is an equilibrium and, by 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), $u_i(s^* + 1) > u_i(s^*)$. This is a contradiction to the assumption that s^* is the most efficient interior equilibrium. ■

Lemma 7 *If $s^e \doteq (s + \overline{k})$ where $\overline{k} = \arg max_{k \in Z} u(s + k)$ and $s_i = s_j$ (s^e is Pareto-best along the diagonal) then there is no (s_i', s_j') such that $u_i(s') > u_i(s^e)$ for $i = 1, 2$.*

Proof. Let WLOG $s_j^{'} < s_i'$ and $s_i' - s_i^e = k$. Then $u_i(s^e) > u_i(s^e + k) = u_i(s_i', s_i') > u_i(s_i', s_j')$ since the payoff is increasing in the action of the opponent. Thus s^e is efficient. ■

Lemma 8 *Let s^* be a symmetric interior UG equilibrium. If $s^* + \overline{k}$ is efficient, then $\overline{k} > 0$. There exists $\overline{k}$ such that $s^* + \overline{k}$ is efficient.*

Proof. Let $\phi + \delta \geq 0$. By lemma 6, $\beta_i(s^*) \leq 0$ and $\beta_i(s_i^* + 1, s_j^*) \geq 0$. Now since $\phi + \delta \geq 0$, by lemma 4, $\beta_i(n) \leq 0$ 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 $i = 1, 2$, $u_i(n, n) \geq u_i(s_i, n) > u_i(s_i, s_j)$.

Suppose now that $\phi + \delta < 0$. Let us argue that the profile that maximises each UG payoff along the diagonal, $\max_{k \in Z} u(s^* + k)$ where $s_i^* = s_j^*$, is $s^* + \overline{k}$ for some $\overline{k} > 0$. First let $\sigma + \delta + 2\phi \geq 0$ (iff $u_i(s + k)$ is convex in k). By the third property in lemma 6, $u_i(s^*) - u_i(s^* - k) > 0$ for any $k > 0$. Therefore (n, n) maximises each UG 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 6, $u_i(s^*) > u_i(s^* - 1)$ for $i = 1, 2$. Since the strategy set is bounded, a maximiser $s^* + \overline{k}$ 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 7, the profile that maximises the payoff along the diagonal is efficient. ■

Lemma 9 *Let $f(.,.)$ be convex in each of its arguments and supermodular. Then $f(x + 2, y + 2) - 2f(x + 1, y + 1) + f(x, y) \geq 0$*

Proof. Let f be convex and supermodular. Then

$$
\begin{aligned}
& f(x+2,y+2) - f(x+1,y+1) - [f(x+1,y+1) - f(x,y)] \\
= \; & f(x,y) - f(x+1,y) - f(x+1,y) + f(x+2,y) \\
& + f(x+2,y+2) - f(x+2,y+1) - f(x+2,y+1) + f(x+2,y) \\
& + f(x+2,y+1) - f(x+2,y) - f(x+1,y+1) + f(x+1,y) \\
& + f(x+2,y+1) - f(x+2,y) - f(x+1,y+1) + f(x+1,y) \\
\geq \; & 0
\end{aligned}
$$

The first effect on the RHS of the equality is the second order effect of the first variable, the second row is the second order effect of the second variable and the remaining two rows are identical and equal to the cross (supermodularity) effect. ■

Lemma 10 *Let $\phi+\delta<0$, $2\phi+\delta+\sigma\geq 0$. Let s be such that $s_i=s_j$. Let $u_i(s)-u_i(s-1)\geq 0$. Let g satisfy {5}. Suppose that $\beta_i(s-1)\geq\theta_i g(v_i(s-1),\eta_j(s-1))$. If $\beta_i(s)\leq\theta_i g(v_i(s),\eta_j(s))$ then $\beta_i(s+k)\leq\theta_i g(v_i(s+k),\eta_j(s+k))$ for all $k>0$.*

Proof. $\delta+2\phi+\sigma\geq 0$ and $\phi+\delta<0$ implies that $\phi+\sigma\geq 0$. Then, by lemma 4, $\beta_i(s+k)$ is increasing and concave in k and $\eta_j(s+k)$ 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, $g(v_i(s+k),\eta_j(s))$ is convex and non-decreasing in k since g is convex and non-decreasing in v by $\{5\}$. Similarly, $g(v_i(s),\eta_j(s+k))$ is convex and non-decreasing in k since g is convex in η for $\eta\geq 0$ by $\{4\}$.

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

$$\begin{aligned} & \beta_i(s)-\beta_i(s-1) \\ \leq\ & \theta_i g(v_i(s),\eta_j(s))-\theta_i g(v_i(s-1),\eta_j(s-1)) \end{aligned}$$

Thus, by lemma 9 and since g is supermodular and convex in its arguments

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

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

Lemma 11 *Let $\phi+\delta<0$, $\phi+\sigma\geq 0$. Let $g(v',\eta)=g(v,\eta)$ for all η and $v',v>0$. Let s be such that $s_i=s_j$. Suppose that $\beta_i(s-1)\geq\theta_i g(v_i(s-1),\eta_j(s-1))$. If $\beta_i(s)\leq\theta_i g(v_i(s),\eta_j(s))$ then $\beta_i(s+k)\leq\theta_i g(v_i(s+k),\eta_j(s+k))$ for all $k>0$.*

Proof. By lemma 4, $\beta_i(s+k)$ is increasing and concave in k and $\eta_j(s+k)$ is non-decreasing and convex in k.

Also $g(v_i(s),\eta_j(s+k))$ is convex and non-decreasing in k since g is convex in η for $\eta\geq 0$ by $\{4\}$, and for all $\widetilde{v}>0$, $g(\widetilde{v},\eta_j(s+k))=g(v_i(s+k),\eta_j(s+k))$ by assumption.

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

$$\begin{aligned} & \beta_i(s) - \beta_i(s-1) \\ \leq \; & \theta_i g(v_i(s), \eta_j(s)) - \theta_i g(v_i(s-1), \eta_j(s-1)) \end{aligned}$$

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

$$\begin{aligned} 0 \; &\leq \; \beta_i(s+1) - \beta_i(s) \\ &= \; -\delta - \phi \\ &= \; \beta_i(s) - \beta_i(s-1) \\ &\leq \; \theta_i g(v_i(s), \eta_j(s)) - \theta_i g(v_i(s-1), \eta_j(s-1)) \\ &= \; \theta_i g(v_i(s+1), \eta_j(s)) - \theta_i g(v_i(s), \eta_j(s-1)) \\ &\leq \; \theta_i g(v_i(s+1), \eta_j(s+1)) - \theta_i g(v_i(s), \eta_j(s)) \end{aligned}$$

We can proceed by induction to show that for every $s+k$ with $k > 0$, we have $\beta_i(s+k) - \theta_i g(v_i(s+k), \eta_j(s+k)) \leq \beta_i(s) - \theta_i g(v_i(s), \eta_j(s)) \leq 0$. Thus every $s+k$ with $k > 0$ is agreeable. ■

Proof of the theorem

Let $\phi + \delta \geq 0$. By lemma 6, $\beta_i(s^*) \leq 0$ and $\beta_i(s_i^* + 1, s_j^*) \geq 0$. Now since $\phi + \delta \geq 0$ by lemma 4, $\beta_i(n) \leq 0$ 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 4, (n, n) is agreeable since (n, n) is an UG equilibrium.

Let $\phi + \delta < 0$ hold. By assumption for each $i = 1, 2$ there are k_i' and k_i'' s.t. $0 \leq k_i'' < k_i'$, $\beta_i(s^* + k_i'') \geq \theta_i g(v_i(s^* + k_i''), \eta_j(s^* + k_i''))$ and $\beta_i(s^* + k_i') \leq \theta_i g(v_i(s^* + k_i'), \eta_j(s^* + k_i'))$. Thus, if $u_i(s^* + k)$ is convex in k (iff $\sigma + \delta + 2\phi \geq 0$) and g satisfies $\{5\}$ we can apply lemma 10. On the other hand, if the marginal harm, $\eta_j(s^* + k)$, is non-decreasing in k (iff $\phi + \sigma \geq 0$) and $g(v', \eta) = g(v, \eta)$ for all η and $v', v > 0$ we can apply lemma 11. In the first case $s^* + \overline{k} = (n, n)$ and thus necessarily $\overline{k} \geq k_i'$ for $i = 1, 2$. In the second case, by assumption $k_i' < \overline{k}$ for $i = 1, 2$. Thus in either case, $s^* + \overline{k}$ is agreeable.■

Proof of the corollary

If $\delta + \phi \geq 0$, since $\beta_i(s^*) = 0$, (n, n) is an equilibrium and we have a contradiction. If $\delta + \phi < 0$ no symmetric action profile where $s_i \leq s_i^*$ is agreeable. To see this notice that, since $\beta_i(s^*) = 0$ and since lemma 4 holds, $\beta_i(s) < 0$. Therefore if $\beta_i(s_i + 1, s_j) \geq 0$, s is an equilibrium contradicting the uniqueness of equilibria. Thus $\beta_i(s_i + 1, s_j) < 0$ implying that $s \notin M^F$. Therefore there is $k_i' > 0$ such that $s^* + k_i'$ is agreeable for i implying $\beta_i(s^* + k_i') \leq \theta_i g(v_i(s^* + k_i'), \eta_j(s^* + k_i'))$. On the other hand s^* satifises $\beta_i(s^*) \geq \theta_i g(0, \eta_j(s^*)) = 0$ as the unique UG equilibrium. Thus, the claim follows from the theorem.■

References

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

[2] Aumann R. (1990): Nash Equilibria Are Not Self-enforcing. In Gaszewitz, Richard, Wolsey: Economic Decision Making, Games, Econometrics and Optimisation p.201-206. Elsevier. Amsterdam, Holland.

[3] Battigalli, P.; Dufwenberg, M. (2006): Dynamic Psychological Games. Mimeo. Bocconi University, Milan.

[4] Baumeister, R.F.; Stillwell, A.M.; Heatherton, T.F. (1994): Guilt: An interpersonal Approach. Psychological Bulletin. 115, 243-267

[5] Baumeister, R.F.; Stillwell, A.M.; Heatherton, T.F. (1995): Guilt as Interpersonal Phenomenon: Two Studies Using Autobiographical Narratives. In Self-conscious Emotions: Shame, Guilt, Embarrassment, and Pride. J.P. Tangney and K.W. Fischer (Eds.). New York: Guilford Press.

[6] Baumeister, R.F.; Stillwell, A.M.; Heatherton, T.F. (1995): Personal Narratives About Guilt: Role in Action Control and Interpersonal Relationships. Basic and Applied Social Psychology. 17, 173-198.

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

[8] Brosig, J.;Weimann J.; Ockenfels, A. (2003): The Effect of Communication Media on Cooperation. German Economic Review 4, 217-242.

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

[10] Charness, G.; Dufwenberg M. (2003): Promises & Parnership. Stockholm University. Working Paper 3/03.

[11] Charness, G.; Dufwenberg M. (2006): Promises & Partnership. Econometrica, forthcoming.

[12] Clark, M.S. (1984): Record Keeping in Two Types of Relationships. Journal of Personality and Social Psychology 47, 549-557.

[13] Clark, M. S.; Mills, J. (1979): Interpersonal Attraction in Exchange and Communal Relationships. Journal of personality and social psychology 37, 12-24.

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

[15] Dawes R.; McTavish J.; Shaklee H. (1977): Journal of Personality and Social Psychology 35, 1-11.

[16] Dawes R.; Orbell J.; van de Kragt A. (1990): The Limits of Multilateral Promising. Ethics, 100, 616-627.

[17] Duffy, J.; Feltowich N. (2002): Do Actions Speak Louder than Words? An Experimental Comparison of Observation and Cheap Talk. Games and Economic Behavior 39: 1-27.

[18] Duffy, J.; Feltowich N. (2006): Words, Deeds and Lies: Strategic Behavior in Games with Multiple Signals. Review of Economic Studies 73, 669-688.

[19] Dufwenberg, M. (2002): Marital Investments, Time Consistency and Emotions. Journal of Economic Behavior & Organization 48, 57-69.

[20] Dufwenberg, M; Kirchsteiger G. (2004): A Theory of Sequential Reciprocity. Games and Economic Behavior 47, 268–298.

[21] Ellingsen, T. ; Johanneson, M. (2004): Promises, Threats, and Fairness. Economic Journal 114, 397-420.

[22] Elster, J. (1989), Social Norms and Economic Theory, Journal of Economic Perspectives. 3, 99-117.

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

[24] Farrell, J.; Rabin M.(1996): Cheap Talk. Journal of Economic Perspectives. 10, 103-118.

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

[26] Frank R.H. (1988): Passions within Reason: The Strategic Role of Emotions. Norton. NY.

[27] Geanokoplos, J.; Pearce D. ; Stachetti, E. (1989) Psychological Games and Sequential Rationality. Games and Economic Behaviour 1, 60-79.

[28] Gneezy, U. (2005): Deception: The Role of Consequences. American Economic Review 95, 384-394.

[29] Hoffman, M.L. (1982): Development of Prosocial Motivation: Empathy and Guilt. In the development of prosocial behavior. N. Eisenberg (Ed.). San Diego, CA: Academic Press.

[30] Huck, S.; Kubler, D.; Weibull J. (2003): Social Norms and Economic Incentives in Firms. ELSE Working Paper. University College London.

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

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

[33] Ledyard, J.O. (1995): Public Goods: A Survey of Experimental Research. In the Handbook of Experimental Economics. J.H. Kagel & A. Roth (eds.). Princeton University Press, Princeton, NJ.

[34] Lev-on, A. (2005): Computer-Mediated Communication and Cooperation in Social Dilemmas: An Experimental Analysis. Mimeo. University of Pennsylvania.

[35] Loomis, J. (1959): Communication: The Development of Trust and Cooperative Behavior. Human Relations 12, 305-315.

[36] Miettinen, T. (2006a): Promises and Conventions - A Theory of Pre-play Agreements. Discussion Paper No 97 / February 2006. Helsinki Center of Economic Research, Discussion Papers,

[37] Miettinen, T. (2006b): Pre-play Negotiations, Learning and Nash-Equilibrium. PhD Thesis. University College London.

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

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

[40] Rabin, M. (1993): Incorporating Fairness into Game Theory and Economics. American Economic Review, 82, 1281-1302

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

[42] Radlow, R; Weidner, M. (1966): Unforced Commitments in 'Cooperative' and 'Non-cooperative' Non-constant-sum Games. Journal of Conflict Resolution 10, 497-505.

[43] Rawls, J. (1972): A Theory of Justice. Oxford University Press, Oxford.

[44] Suetens, S. (2005): Cooperative and Noncooperative R&D in Experimental Duopoly Markets. International Journal of Industrial Organization.

[45] Suetens, S.; Potters, J. (2006): Bertrand Colludes More Than Cournot. Experimental Economics, forthcoming.

[46] Smith A. (1759): The Theory of Moral Sentiments. Reprinted in (2002). Ed. Knud Haakonsen. Cambridge University Press. Cambridge, UK.